ED 367 147 FL 021 869

AUTHOR	Anivan, Sarinee, Ed.
TITLE	Issues in Language Programme Evaluation in the 1990s. Anthology Series 27.
INSTITUTION	Southeast Asian Ministers of Education Organization (Singapore). Regional Language Centre.
REPORT NO	ISBN-9971-74-039-7; ISSN-0129-8895; RELC-P385-91
PUB DATE	91
NOTE	174p.; Selected papers from the Southeast Asian Ministers of Education Organization (SEAMEO) Regional Language Centre Seminar (Singapore, 1990). For individual papers, see FL 021 870-882.
PUB TYPE	Collected Works - General (020) -- Collected Works - Serials (022)
EDRS PRICE	MF01/PC07 Plus Postage.
DESCRIPTORS	Curriculum Evaluation; Elementary Secondary Education; *English (Second Language); English for Academic Purposes; Evaluation Criteria; Evaluation Methods; Foreign Countries; Formative Evaluation; Higher Education; Intensive Language Courses; Language Proficiency; Language Role; Language Teachers; *Language Tests; *Program Evaluation; Research Utilization; Second Language Instruction; *Second Language Programs; Self Evaluation (Individuals); Teacher Education; Teaching Methods; Testing; Theory Practice Relationship; Writing Instruction
IDENTIFIERS	Indonesia; Japan; Malaysia; Nepal; Singapore

ABSTRACT

Papers on aspects of second language program evaluation include the following: "The Role of Language Testing in Language Program Evaluation" (Adrian Palmer); "Defining an Appropriate Role for Language Tests in Intensive English Language Programs" (Margaret Des Brisay, Doreen Ready); "Principles and Practice in an Evaluation Project" (Dermot F. Murphy); "National Level Formative Evaluation: Some First Steps" (Ali Abdul Ghani, Brian Hunt); "Second Language Proficiency Assessment and Program Evaluation" (David Nunan); "How Program Personnel Can Help Maximize the Utility of Language Program Evaluations" (Ronald Mackay); "The Development of Self Assessment Skills in TESOL Teacher Preparation" (Alastair L. McGregor); "Bringing Evaluation and Methodology Closer Together" (David Crabbe); "Evaluating a Teacher Training Project in Difficult Circumstances" (C. J. Weir, J. Roberts); "Program Evaluation in Light of Language Learning Background, Student Assessments, and TOEFL Performance" (Harry L. Gradman, Edith Hanania); "Towards Evaluating the Writing Laboratory: A Prototype" (Ma Flor E. Mejorada, Elvira Fonacier); "Trialling of the NEW EL Syllabuses for Singapore Schools" (Goh Soon Guan); and "EAP Program Evaluation in an Asian Context: A Case from Japan" (Mark Sawyer). (MSE)

Issues in Language Programme Evaluation in the 1990s

Edited by
Sarinee Anivan

ANTHOLOGY SERIES 2

ISSUES IN LANGUAGE PROGRAMME EVALUATION IN THE 1990s

ISSUES IN LANGUAGE PROGRAMME EVALUATION IN THE 1990s

Edited by Sarinee Anivan

Anthology Series 27
Published by
SEAMEO Regional Language Centre

30 Orange Grove Road
Singapore 1025
ISBN 9971-74-039-7
ISSN 0129-8895

Note: The views expressed by the authors are their own and do not necessarily represent those of RELC

CONTENTS

FOREWORD

The theme of the 1990 RELC Regional Seminar was "Language Testing and Language Programme Evaluation". The present volume contains a selection of papers on language programme evaluation from the Seminar, while selected papers on language testing are presented in a separate volume.

Language educators need access to worthwhile studies in the area of programme evaluation, both case studies and reports of individual evaluation projects, and theoretical and methodological discussions. Yet the relevant documents sometimes do not become widely available because of confidentiality requirements. When evaluation studies are published, they are sometimes too specific to be of interest to those not immediately involved with a particular project.

Such problems and limitations make all the more valuable a collection of papers such as the present one which is, I believe, the first ever RELC book-length publication devoted entirely to language programme evaluation. It contains papers by a number of well-known experts in the field and their studies will, I feel sure, enrich our understanding and provide models and insights for future work.

I am pleased to note that some of the papers deal with language education programmes in Southeast Asia. In this dynamic and fast-developing region, the provision of education is a vast enterprise requiring the investment of very considerable resources, both human and financial. It is also an enterprise that reflects and focuses the hopes and aspirations of millions of people. In view of this, the importance of effective programme evaluation must be obvious. Society has the right to know that the teaching-learning process is carefully and periodically reviewed. Evaluation programmes, if sensitively carried out, also take into account the needs and values of a society, concerns that are reflected in the present volume.

I commend this anthology to the attention of language educators both within and outside Southeast Asia. I am confident they will find it a useful and timely addition to the literature on language teaching programme evaluation.

Earnest Lau
Director, RELC
January 1991

INTRODUCTION

The theme of the 1990 RELC Regional Seminar, "Language Testing and Language Programme Evaluation", acknowledges the vital link between the two aspects of language education. Test results can only make a worthwhile contribution if tests are designed properly to answer questions raised in evaluation. **Palmer's** analysis of eight method comparison studies shows that test results cannot always say definitively which is the more effective method when there are limitations in the testing procedures. For example, comparisons would be on a stronger basis if programme-free, criterion-referenced tests were included.

Nunan also advocates more use of criterion-referenced testing in programme evaluation. Given the difficulties in the definition and measurement of the criteria of general language proficiency, he proposes using some curriculum-bound version of criterion-referenced testing. Similarly, **Des Brisay and Ready** warn against using gain scores as a final measure of the effectiveness of a programme. They report instances of wide unexplained variation in TOEFL scores of some individual students. Their discussion of moves to improve tests, and set up programmes which rely less on test results in deciding on trainees' potential for further studies, are of interest to the Southeast Asian region.

According to **Murphy**, change and innovation in language programmes should be accompanied by well planned formative evaluation from the earliest stages to avoid costly mistakes. Innovation would have a better chance of success if supported by the lower level personnel who have to implement change.

While Murphy describes in broad outline the evaluation of the new English curriculum in Malaysia, **Ghani and Hunt** provide greater details of the preparations leading to that evaluation project. **Goh** reports on the evaluation of trials of a new English Language syllabus in Singapore schools.

Mackay stresses the importance of inputs from stakeholders such as teachers and administrators, and suggests ways for such people to get a hearing at the earliest stage. Their input would thus have a better chance for consideration and inclusion, which would result in better evaluations.

Crabbe proposes methodology that gives learners a part in the evaluation of their own performance, and also of the effectiveness of the programme, while taking a course. He argues that in this way better learning strategies can evolve, which will help learners to attain a higher level of self-direction in their learning.

When looking at test scores to estimate the effectiveness of programmes, the factors that appear to contribute to the scores should also be explored. **Gradman and Hanania** researched how forty-four learning background factors correlated with TOEFL scores. For example, they found "extensive outside reading" and teacher quality to be significant factors.

If the teacher is a factor in language learning, the quality of teacher training programmes should be of interest to evaluators. **McGregor** looks at the development of self-assessment skills in language teacher training programmes. With such skills, teachers can improve and adapt to change. Self awareness can lead to awareness of and concern for the needs of learners. Some of the problems encountered by **Weir and Roberts** when evaluating a teacher training programme in Nepal can perhaps be lessons to others involved in similar projects in "difficult circumstances", when compromise and improvisation are necessary.

Studies on a smaller scale are also important forms of evaluation. **Mejorada** and **Fonacier** describe the evaluation of a writing programme in the Philippines, while Sawyer gives details of the evaluation of an EAP programme in Japan. But whatever the scale, evaluators are looking for higher effectiveness, including cost effectiveness, and appropriateness of programmes. Test scores are perhaps the most important data used in making evaluation judgements. But they are not the only factors to be considered. Almost all the papers in this volume define evaluation in a broader sense to include the assessment of any part of a language programme.

In conclusion one can say that the range of difficulties encountered in the design and implementation of language programme evaluation is as wide as the range of socio-economic situations among Southeast Asian countries. However, we hope that each study presented here will contain at least some generalisable ideas.

Sarinee Anivan

THE ROLE OF LANGUAGE TESTING IN LANGUAGE PROGRAM EVALUATION

Adrian Palmer

INTRODUCTION

First, let me say how happy I am to be here.* When I started to prepare for this talk, I recalled a paper Jack Upshur, my mentor in Language Testing presented at the SEAMEO Regional Language Center Seminar on Language Testing exactly twenty years ago. I remembered that the copy was made on a thermofax machine in pre-Xerox days, and I realized just how long I have been influenced by RELC. To be honest, in addition to my professional interest in language testing and program evaluation, I also value the opportunity to renew old friendships. This is why coming to a RELC seminar is a double treat for me.

What I plan to do today is focus on the interpretability of test scores in program evaluation studies. I will examine two main issues: test design issues and research design issues.

First, I will briefly describe eight method-comparison, program evaluation (MC-PE) studies comparing acquisition-based and analysis/practice based methods. (Since analysis/practice is both somewhat clumsy and also perhaps overly limiting, I will use the terms "traditional" or "eclectic" to refer to this method, even though these terms do not describe the basis of the method in the same way that "acquisition" does for the experimental method.)

Second, I will describe two test-design issues: the basis for the tests (syllabus vs. theory) and the choice of scales (norm referenced vs. criterion referenced). After introducing the issues, I will analyze the choices made in each of the studies to illustrate how these considerations were dealt with in actual research.

Third, I will describe two research-design issues: instructional purity and subject selection, and I will analyze the choices made in each of the studies.

Finally, I will summarize and analyze the results of the studies and suggest areas of test development that require our attention.

Fourth, I will present the results of a comparative analysis of the outcomes of the studies.

Finally, I will discuss some of the assumptions which must be made in order for the conclusions about overall results to be valid and make some suggestions for future directions in language test development and use.

THE METHODS

The eight program evaluation studies have a common theme: comparison between acquisition-based language teaching methods and traditional methods. I will define acquisition methods as those that expose the student to the language as a whole, anticipating that the student will pick up the structure, etc., subconsciously. Traditional methods are those that also use analysis, practice, and explanation in order to build overall competence.

* I would like to thank Paul Kramer for our many years of collaboration on issues related to testing the Input Hypothesis, as well as some of the specific analyses presented here. I would also like to thank Lyle Bachman for our discussions of interpretability of test scores and research results, as well as his specific comments on this paper.

In this paper, I will focus almost entirely upon a comparison of the effectiveness of acquisition based versus traditional instruction in promoting the development of general language proficiency. I would emphasize, however, that the individual studies also looked at program-specific outcomes (such as academic subject-matter learning).

The specific theory of language acquisition upon which the most of the studies were based is Stephen Krashen's Input Hypothesis (Krashen, 1985). In its strongest form, this theory states that two and only two factors are responsible for second language acquisition: comprehensible input and low affective filter strength.

The eight studies reviewed include three studies of content-based, second language instruction (Burger, 1989; Edwards, Wesche, Krashen, Clement, & Kruidenier, 1984; Hauptman, Wesche, and Ready,1988); two studies of content-based foreign language instruction (Lafayette & Buscaglia, 1985; Sternfeld, 1989); and three studies of non-content based, foreign language instruction (Asher, Kusudo, & de la Torre, 1983; Lightbown, 1989; Kramer, 1989).

TEST DESIGN ISSUES

The first issue I will investigate is the relationship between the test designs used in the studies and their interpretability. I will focus on four major test design options, involving two issues: the test content and the kinds of scales used. These options are outlined below in Figure 1.

Figure 1

Test Design Options

	Test Content	
Scales	Proficiency	Achievement
Norm referenced		
Criterion referenced		

TEST CONTENT: ACHIEVEMENT VS. PROFICIENCY

Basic Considerations

When developing or choosing tests for program evaluation, one of the first questions that arises is what to test. Beretta (1986a) describes three design patterns for testing: program specific achievement tests for each program, program-neutral proficiency tests, and a combination of achievement tests program-specific plus program-neutral measures.

The content of achievement tests is based upon a syllabus and samples what the students were taught. The strength of achievement tests is that one does not have to defend the course objectives. One has only to demonstrate that the tests cover a reasonable sample of the material taught. The weakness of achievement tests in MC-PE studies is that comparisons must be made at least partially in terms of the programs' effectiveness in covering material they were not designed to cover.

Proficiency measures are based upon a program-neutral theory of language and provide a way of directly comparing the relative effectiveness of different programs in reaching program neutral goals (but don't provide a means of evaluating the effectiveness of different programs in reaching their own specific goals). They allow us to ask the question "what is the relative effectiveness of these two programs in accomplishing thus and so?" Using proficiency tests requires us to address two major questions: What is the nature of the language competence, and what evidence do we have that the tests we are using actually measure that competence?

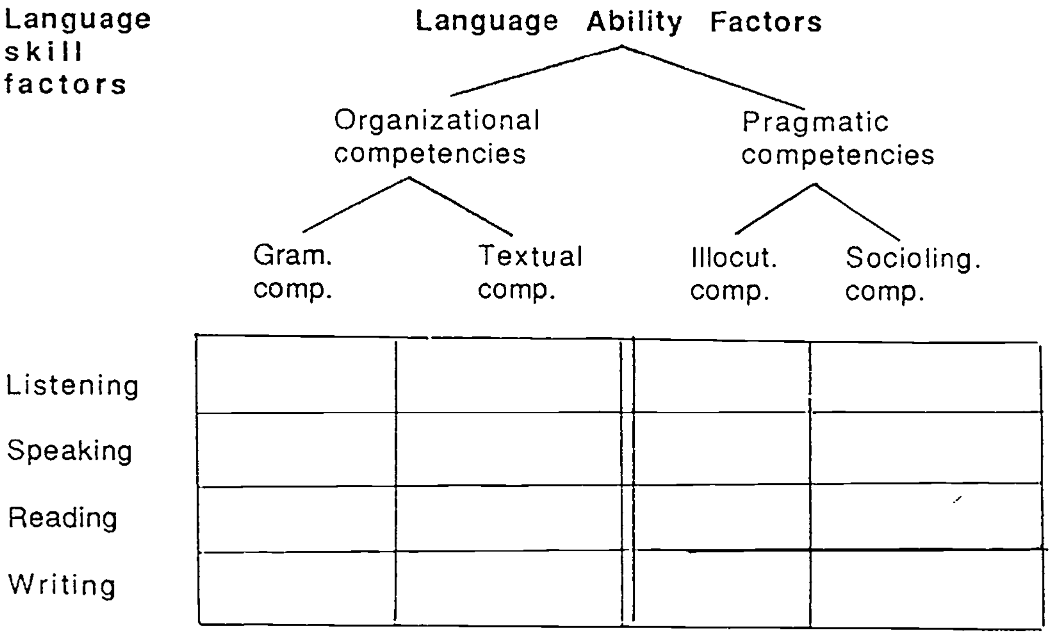
MODELS OF LANGUAGE ABILITY

The two models of language ability which seem to have attracted the most interest in the past decade are those inspired by Canale & Swain (C-S), and those inspired by Oller. These two models contrast in that the C-S model attempts to describe the various components of language ability, while the Oller model focuses primarily on the communality.

Lyle Bachman and I have worked extensively with the Canale-Swain model, and we have adopted a version which I will call the organizational-pragmatic model. In addition to these two major constructs, the model also includes the four language use skills of listening, reading, speaking, an writing.

Figure 2

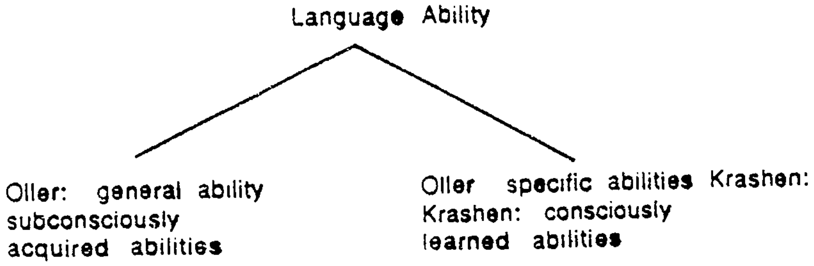
Communicative Language Ability Constructs



The Oller inspired model(s) include two major constructs: a large general ability construct, and some smaller specific constructs. Krashen has sometimes interpreted these two constructs as "acquired" and "learned" competencies, an interpretation which I find reasonably compatible with Oller's.

Fig. 3

Oller (Krashen?) Model of
Language Ability


Language Ability
Oller: general ability subconsciously acquired abilities
Oller specific abilities Krashen: Krashen: consciously learned abilities

EVIDENCE OF CONSTRUCT VALIDITY

Over the past decade and longer, researchers have been devoted considerable effort toward investigating the construct validity of these models. Here are four general conclusions which I believe the research supports. First, there is a distinction among the language use skills (listening, speaking, reading, and writing). Second there is a distinction between organizational and pragmatic competencies. Third, in addition to distinct abilities, a general ability factor affects all language test scores. And finally, language test scores are affected by test method (such as multiple-choice, cloze procedure, translation procedure, interactive interview, self-rating procedures, etc.) The evidence supporting these generalizations is found in a growing body of research, including Bachman 1982, Bachman & Palmer, 1981, 1982, 1989; Brütsch, 1979; Clifford, 1981; Fouly, Bachman, & Cziko, 1990; Oller 1979 & 1983; Palmer, 1972; Upshur & Homburg, 1983; and Upshur & Palmer, 1974.

The point of this is to emphasize that language testing researchers have been thinking about the nature of communicative language ability for some time and have been developing and evaluating proficiency tests based upon recent models of language ability.

ANALYSIS OF ACHIEVEMENT/PROFICIENCY CONSIDERATIONS

Given this relatively large body of language testing research on construct validity, it is interesting to examine the language tests used in MC-PE studies to investigate the extent to which they have been influenced by these developments in language testing research.

Table 1

Analysis of Tests Used in MC-PE Studies

Study	Specificity of Measures		Scoring Reference		Language Ability Model			
	Ach.	Prof.	NR	CR	Skls.	Meth.	Gen-spec./ Acq.-Learn	C-S B-P
CB-SL								
Burger		+	+		+	+		
Edwards et al.		+	+		+	+		
Hauptman et al.		+	+		+	+		
CB-FL								
Lafayette et al		+	+		+			
Sternfeld		+	+		+	+		
NON-CB, FL								
Asher et al.	+	+	+		+			
Kramer		+	+		+	+	+	
Lightbown	+	+	+		+			

As can be seen from the first two columns in Table 1, two of the eight studies reviewed (Asher et al. and Lightbown) included syllabus-based achievement tests. All eight of the studies used proficiency measures reflecting distinctions among the language skills constructs (listening, speaking, reading, and writing), although not all studies used tests of all four skills.

In addition, as can be seen in the four columns under "Language Ability Model," about half of the studies also used proficiency tests classified by test method (translation, cloze, summary, and multiple choice. None of the tests used seems to have been directly influenced by the Canale-Swain/Bachman-Palmer model of communicative language ability. However, the use of cloze testing procedures in several of the studies does suggest the influence of Oller's work on general vs. specific factors, and possibly (by extension) Krashen's acquired/learned competence constructs.

In most of the studies, the tests used are named, and in some cases described, but not systematically classified by trait and method. Under Instruments, for example, we might find a list of tests such as "reading," "vocabulary" "grammar," "cloze," and "translation." Notice that such a list classifies tests sometimes by language use skill, sometimes by language ability, and sometimes by method. What we do not find are descriptions of the theory of language abilities upon which the tests were based. Nor do we find consistent distinctions made between language ability and test method.

In addition, we find almost no references to the specific language ability constructs which form the heart of Krashen's input hypothesis: acquired and learned competencies. And while many of the studies include tests commonly thought of as "integrative" (such as cloze and dictation) and "discrete-point," (such as multiple-choice grammar), reference is generally not made to the possible relationship between such tests and the primary language ability constructs in Krashen's theory.

One exception to this is Kramer, who provided a lengthy discussion of issues involved addressing the construct validity of the measures used. Analyzing the pattern of scores on his tests, he discussed the validity of the measures in terms of the basic constructs in Krashen's Input Hypothesis: "acquired:" and "learned" competence." While Kramer was not able to employ measures with prior demonstrated construct validity, because of the purity of his instruction (see below) he was able to assess whether the results of the research provided any evidence for the validity of the acquired/learned competence distinction.

In summary, with respect to the issue of the construct validity of the language tests used in methods comparison program evaluation studies, I believe what we see here is a general trend for such studies to employ tests that might be considered deficient in the following ways. They lag behind recent work in language testing research; they use tests which are based upon models different from those that the methods' developers had in mind when they developed their methods; and they use tests which tend to avoid the issue of the distinction between language trait and testing method.

SCALING ISSUES

Overview

I now turn to the choice of frame of reference for interpreting test scores. Norm-referenced (NR) scores are interpreted only in reference to the performance of a particular group of individuals. In contrast, criterion-referenced (CR) tests scores are "...interpreted as an indication of an individual's attainment with respect to a given domain of proficiency" (Bachman & Clark 1987:28). Each frame of reference has its own strengths and weaknesses (see Brown, 1989).

One of the main reason that norm referenced are so widely used is that they are available. And probably one reason they are so available is that they are easy to construct. We can get away without defining what it is that we are measuring at all! People are compared to people, not to levels of ability, which means that the nature of language ability can go unspecified. This factor which contributes to their ease of construction is also, of course, one of their main weaknesses. Another weakness of NR scales is that they do not provide us with a measure of how much of a body of knowledge one controls. Thus, they do not provide us with the kinds of information we might want in assessing the relative importance of attained levels of ability.

The main strength of criterion referenced scores of language ability is that ratings are comparable across a wide range of contexts and content areas (Bachman & Savignon, 1985), which is precisely what Bachman (1989) suggests would be useful in MC-PE studies. Criterion referenced scores would allow us to address very interesting issues, such as the amount of given ability that students have mastered. This, in turn, would allow us to assess the importance of that level of mastery.

The main weakness of these scales is the practical difficulty in constructing them for tests of general language ability. Specifically, one needs to define language ability precisely, to keep as distinct as possible the roles of language ability and test method, to keep distinct language ability and context, and to specify zero and complete mastery levels.

Bachman and I tried to define such scales in our work, and Bachman and Clark (1987) have suggested a general program to further develop, refine, and operationalize such scales. Much work remains to be done, both on the conceptual and operational level, before we have available a battery of CR language ability measures, but, as I hope to show, given the kinds of outcomes we are getting in our much of our MC-PE research, and given the difficulties presented in interpreting these outcomes, I think this kind of research and development work is warranted.

ANALYSIS OF THE SCALE OPTIONS USED IN THE STUDIES

All of the studies reviewed used NR scales (see "Scoring Reference" in Table 1). Thus, while we can say that one group of students performed better than another group, we cannot say how much of the language either group controlled. This means that we cannot reach conclusions about the importance of the levels of competence reached, nor about the relative effectiveness of each program in reaching its own unique objectives.

Some of the MC-PE researchers have noted this problem and provided additional information to make the results more interpretable. Both Kramer and Sternfeld provided examples of what the students were able to do after completing the program of study. Kramer also provided descriptions of student performance.

The need to go beyond the typical NR comparative statistics became particularly obvious to me when I attended a meeting between a dean at the University of Utah and a group of researchers. The Dean took one look at the summary statistics and immediately said, "Setting the differences between groups aside, just how much of the language have these people learned in one year of instruction?" The Dean was more concerned with the amount the students had learned than with what appeared to be minor (though possibly statistically significant) differences between groups.

RESEARCH DESIGN ISSUES

So far, I have examined the effect of test design on interpretation of the results of the research. Now, I turn to the influence of the research designs and their effect on interpretation of test scores. Specifically, I will examine two design issues: the purity of the instruction used and the backgrounds of the subjects involved in the studies. While these are by no means all of the relevant internal validity considerations, they strike me as being particularly important in MC-PE research.

INSTRUCTION PURITY

Overview

Instructional purity is the extent to which the treatments of the two groups of subjects are faithful to the theories on which they are based. Instructional purity affects the internal validity (Beretta 1986b, Brown 1988) of the study, which is the extent to which we can attribute results (as measured by tests) to differences in treatment. Studies in which such attributions can be logically made are said to exhibit a high degree of internal validity.

If we wish to compare the effectiveness methods based upon different principles, such as acquisition versus analysis/practice we need to start out with definitions of the methods based upon the principles, "descriptive data" (Richards & Rodgers, 1986: 181-3). According to Krashen's definitions, language acquisition is the result of comprehensible input and low filter strength, and nothing else. This indicates what must be provided in "acquisition" classes. In contrast, "traditional" or "eclectic" instruction might be operationalized as instruction which also provided conscious learning, drills, production oriented activities, practise, explanation, analysis, etc. (along with comprehensible input).

Once we have defined the theoretical bases for and differences between the methods being compared, we need to look for some sort of evidence (observational data) that the activities that took place in the classroom were faithful to these definitions and distinctions.

ANALYSIS OF THE STUDIES WITH RESPECT TO PURITY OF INSTRUCTION

Paul Kramer and I analyzed the reports of the eight MC studies to determine the extent to which conscious learning and production activities were included in the experimental (acquisition based) classes, these being the primary activities which are said to contribute to learning, but not to acquisition (Kramer and Palmer, 1990). The results of our analysis are given in the first two columns of Table 2.

Table 2

Analysis of Factors
Affecting the Internal Validity of the Studies

	Learning Activities		Stud. Background
Study	Consc. Learn.	Production	(Prior Instruction)
CB-SL			
Burger	yes	yes	advanced
Edwards et al.	no gram. voc (?)	yes	high intermediate
Hauptman et al.	no gram. voc (?)	yes	high intermediate
CB-FL			
Lafayette et al.	no	yes	fourth course
Sternfeld	no	some	some beginning
NON-CB, FL			
Asher et al.	no	yes	beginning (?)
Kramer	no	no	beginning
Lightbown	no	no	beginning

As can be seen in columns 1 and 2 under "Learning Activities" in Table 2, most of the studies seem to have avoided the use of conscious learning activities in the "acquisition" treatment, perhaps with the exception of the conscious teaching of vocabulary. On the other hand, most of the studies seem to have provided situations, such as discussion groups, in which the students were expected to produce the language, as opposed to just processing input. While Krashen's Input Hypothesis certainly does not include a rule prohibiting speech, it does specifically state that speaking is not a cause of language acquisition.

To test Krashen's theory, it seems to me that we must try our best to keep it as distinct as possible from other theories. The two main differences between Krashen's theory and others seems to me to lie in the two negatives: neither conscious learning nor production are required for acquisition. Thus, If we include either of these in the method which is supposed to be an operationalization of Krashen's theory and not other theories, it is difficult to claim that the two methods are distinct.

On the whole, it appears only two of the eight studied (Kramer & Lightbown) provided the students with relatively pure operationalizations of acquisition-based instruction, as defined in Krashen's theory. The research reports generally did not include descriptions of the traditional instruction used for the control groups, but I think it is reasonable to assume that this instruction was eclectic enough to be distinct from the narrower range of activities found in the studies using relatively pure acquisition-based instruction. Nevertheless, the fact that the traditional instruction is not carefully described is a weakness of the reports of these studies.

STUDENTS BACKGROUND

Overview

Another research design factor affecting the interpretation of test scores in MC studies is the nature of the abilities that the students bring with them to the study. If we are comparing the relative effectiveness of two methods of language teaching, and if amount of treatment to which we expose them is small relative to the total amount of prior instruction they have received, and if the testing indicates some sort of positive outcome, it would still be risky to advocate the experimental treatment as the basis for all of a student's language learning activity. It would also be risky to infer that the experimental treatment alone (and the treatment upon which it was based) explained the outcomes obtained.

Students at intermediate or advanced levels of language proficiency might be presumed to have been exposed to a fairly wide range of language learning activities. Adding even a fairly narrowly focussed type of instruction (such as comprehensible input) might have fairly little effect on the overall range and quantity of language learning activities to which these students were exposed; and a method narrowly defined as containing only what was added (such as comprehensible input) would not resemble the total range of instructional activities affecting the results of the research.

Moreover, particularly if you do not employ random assignment to groups, you are likely to run into problems caused by differential backgrounds between the two groups. (This and other design problems are dealt with in some detail in Kramer & Palmer, 1990).

ANALYSIS OF THE STUDIES IN TERMS OF STUDENTS' BACKGROUND

The results of Kramer's and my analysis of the eight studies is given in the third column of Table 2. This indicates that the students were about equally divided between those at the intermediate-advanced level and those who were relative beginners.

So far, I have discussed test and research design and their effects on the interpretation of results, and I have noted that two of the studies (Lightbown's and Kramer's) seem to be more interpretable than others. I now turn to the outcomes of the studies.

QUANTITATIVE RESULTS

When one employs a treatment as radical as a "pure" implementation of acquisition based instruction (no conscious learning, no focus on form, no production activities) with groups of students differing maiкedly in initial language proficiency, as well as age, it is reasonable to hypothesize that there would be significant interaction between treatment and level, or between treatment and age. Such interaction might render invalid any global interpretation of the treatment as "effective" or "ineffective." I will now present a comparative analysis of the results of the studies, which, I believe, illustrate just this sort of interaction.

A between-group comparison of end-of-treatment scores on proficiency tests is given in Table 3 (Kramer & Palmer, 1990).

Table 3

Between-Group
Post-Test Proficiency Comparisons

	TRADITIONAL GROUP SIG. BETTER	NO SIGNIFICANT DIFFERENCES	ACQUISITION GROUP SIG. BETTER
"Pure" Studies	(Adult L2) K-oral int. K-reading trans. K-writing summ. K-vocabulary (4 outcomes)	K x 3 (3 outcomes)	(Child L2) Li-vocabulary Li-pictures Li-speaking* (3 outcomes)
"Impure" Studies	H-cloze S-writing La-reading La-writing (4 outcomes)	H x 11 S x 4 E x 7 La x 1 B x 5 (28 outcomes)	H-translation H-total prof. E-cloze L-speaking (4 outcomes)

NOTES: *Between-group differences on Lightbown's speaking tests were large but were not tested for significance (small N).
No post-test *proficiency* comparisons provided in Asher et al, so no outcomes for this study are included in this table.

The table is constructed to call attention to the interaction between students' age, purity of treatment, and effectiveness of the methods.

Within cells are comparative post-instruction proficiency test outcomes (or gains in those studies employing ANCOVA's with pre-test scores as covariates), designated by the initials of the first author (E = Edwards et al, La = Lafayette, etc.) and test content (speaking, vocabulary, etc.). In the top row are the outcomes for the two relatively "pure" studies. In the bottom row are the data points for the six relatively "impure" studies. In the left column are outcomes significantly favoring traditional treatments. In the center column are outcomes for which no significant post-instructional proficiency differences were found. In the right column are outcomes significantly favoring the experimental (acquisition-based) treatment.

Notations such as "K x 3" (as in the top center cell) indicate that no significant differences were found on three of the post-test proficiency measures in the Kramer's study.

In a few studies, we had to make arbitrary decisions as to whether to include both part and whole test scores as data points, so others who might analyse these studies on their own might arrive at slightly different totals from those than presented here. I believe, however, that the overall trends would likely be the same.

Our first general observation is that in the two "pure" studies, overall effectiveness of instruction was related to the students' age. The students in Kramer's study were adults, while those in Lightbown's study were children. Kramer's MANOVA indicated that the traditional students performed significantly better than the acquisition students both overall and on four of the seven tests. The experimental students in Lightbown's study performed better on all three proficiency measures, significantly better on two of them. Due to the small number of students taking the speaking test, no tests of significance were performed, although the differences appear to be large.

In the "impure" studies, there appears to have been no significant main effect (type of instruction). In addition, most of the comparisons between groups on individual tests indicate no significant differences: 28 non-significant differences versus 8 significant differences (four favoring the traditional group, four favoring the acquisition group). In addition, on those individual tests for which significant differences were found, I do not see any obvious interaction between treatment and specific language ability. For example, on the cloze test in Hauptman et al., the control group outperformed the experimental group; whereas in Edwards et al., the experimental group outperformed the control group.

Normally, when one encounters a large number of non-significant differences (as was the case for the "impure" studies), one would be concerned about the reliability of the measures. With unreliable tests, one would find non-significant between-group differences over and over, and conclusions that one group performed "at least as well" as another group would be meaningless. In addition, statistical logic requires that we first reject the null hypothesis that no learning took place before drawing a conclusion that two groups performed comparably.

Test reliability does not appear to have been a problem in these studies. Most of the studies included evidence of test reliability. And the null hypothesis of no learning has also been addressed, although the fact that some of the studies were conducted in a second-language environment tends to make rejecting the null hypothesis somewhat more problematic.

DISCUSSION

The analysis of the language testing outcomes presented abcve appears to point to fairly straight forward and strong conclusions both about language acquisition theory and about method. It suggests that theories of child and adult language acquisition are different. It suggests that different methods work for children and adults. It suggests that balance.i methods are more effective for adults than methods with a narrow focus. And it suggests that more than comprehensible input (with low filter) is needed for efficient adult L2 acquisition. Yet I would immediately like to caution against taking overly strong positions about the validity of these inferences. Specifically, I would like to point out some of the limitations in our testing procedures which ought to raise caution flags at a number of points.

First, the conclusion that certain programs seem more or less efficient than others in teaching language depends upon our confidence that the tests adequately sample what we believe to be the important components of language ability. If the tests are biased toward one program or another, the results will also be biased. It strikes me that one of the best ways to avoid bias in the selection of program neutral tests of general language ability is to use tests which are clearly related to a theory of language ability, preferably one which has been validated in independent research. If the selection of tests is at all haphazard, to that extent the results of the tests might be expected to be biased.

Second, if we observe significant differences between groups on measures of language ability and make anything of these differences, we are assuming that significant differences are also meaningful differences. As J. D. Brown points out, significance and meaningfulness are two different issues and must be addressed separately (Brown 1988: 141). As long as we continue tc use norm-referenced scoring procedures, I am afraid that we will find it difficult to obtain the kind of information we need to distinguish the significant from important.

Third, because the analysis of the eight studies presented in this report addresses the issue of the program's effectiveness in promoting language acquisition, one might interpret this as a comment on the general value of the programs. This is clearly an unwarranted interpretation. Immediate gains in language proficiency are only one possible measure of a program's value. The researchers, however, were interested in other outcomes as well. For example, in the sheltered subject matter programs, mastery of the subject matter was an important consideration. Also, many universities are interested in promoting area studies programs, and based upon evidence from students' journal entries, students in the University of Utah's acquisition-based programs seem to have become increasingly aware of the L2 culture. What is the relative importance of this outcome compared with the development of language proficiency? And what is the relative importance of affect and attitude, variables measured in many of the studies reviewed?

Another measure of a program's success might be the extent to which its students continued on with additional language study. For example, follow up observation of subjects in Sternfeld's study indicates that a much larger percent of students in the immersion program continued on to more advanced courses. In this case small, (but possibly significant) between-group differences in language ability at the end of one year might prove inconsequential in the long run. If students are a little better than their peers at the end of one year of language study but then quit, within a couple of years this initial difference would be meaningless. After all, length of exposure is an important variable in language acquisition.

Additionally, the apparent clarity of the findings hinge to a considerable degree on two studies: Lightbown's and Kramer's. Just how confident should we be that these findings would be replicated? They might be, but again they might not. Should we be more confident of the patterns observed with a relatively large number of replicated."impure" studies than with a few unreplicated "pure" studies?

CONCLUSIONS

In the 12th annual Language Testing Research Colloquium held in San Francisco, in March of 1990, both Lyle Bachman and I expressed a fear that we as language testing researchers were becoming isolated from other users of language tests, and as a result what we are learning about language and language tests is not having an effect outside of our interest group.

I see the testing components of the eight program-evaluation studies reviewed in this paper as evidence that this fear is justified. I have experienced first hand the practical problems we face in trying to put to use what we have learned in our research. I consulted in both the Kramer and Sternfeld studies and had ample opportunity to influence the testing efforts, yet both of these studies were conducted by under conditions which would have made it difficult to develop or use the kinds of testing designs and procedures being advocated in the field of language testing research. Moreover, individual researchers naturally have their own testing interests and agendas, and who is to say that ours are more important than theirs?

If we as language testers are to have an impact on the use of language tests in program evaluation, or anywhere for that matter, we have to make it practical for others to use what we have discovered. We cannot expect researchers whose interests may be primarily in methodology or theory to find the time and develop the expertise necessary to create new, practical, construct valid criterion-referenced tests of communicative language ability. We need to do this development work on our own and then make tests of this sort available to others.

REFERENCES

Asher, J., Kusudo, J., & de la Torre, R. 1983. Learning a second language through commands: the second field test. In Oller, J. & Richard-Amato, J. Methods That Work: A Smorgasbord of Ideas for Language Learners. Rowley, Mass.: Newbury House. 58-72.

Bachman, L. 1982. The trait structure of cloze test scores. TESOL Quarterly, 16, 61-70.

Bachman, L. 1989. The development and use of criterion-referenced tests of language ability in language program evaluation. In Johnson, R. (ed.). The Second Language Curriculum. Cambridge: Cambridge University Press. 1989.

Bachman, L., and Clark, J. 1987. The measurement of foreign/second language proficiency. ANNALS, AAPSS, 490. March 1987.

Bachman, L., and Palmer, A. 1989. The construct validation of self ratings of communicative language ability. Language Testing, 6.1.

Bachman, L., and Palmer, A. 1982. The construct validation of some components of communicative proficiency. TESOL Quarterly, 16.4. 449-463.

Bachman, L., and Palmer, A. 1981. The construct validation of the FSI Oral Interview. Language Learning, 31.1. 67-86.

Bachman, L., and Savignon, S. 1985. The evaluation of communicative language proficiency: A critique of the ACTFL Oral Interview and suggestions for its revision and development. Paper presented at the "Perspectives on Proficiency" Forum, 1985 Modern Language Association of America Convention, 29 December 1985, Chicago, Ill.

Beretta, Alan. 1986a. Program-fair language teaching evaluation. TESOL Quarterly, 20.3. 431-444.

Beretta, Alan. 1986b. A case for field experimentation in program evaluation. Language Learning, 36.3. 295-309.

Brown, J. D. 1988, Understanding Research in Second Language Learning. Cambridge: Cambridge University Press. 36-40.

Brown, J. D. 1989. Language program evaluation: a synthesis of existing possibilities. In Johnson, R. K. (ed.) The Second Language Curriculum. Cambridge: Cambridge University Press. 222-241.

Brütsch, S. M. 1979. Convergent-discriminant validation of prospective teacher proficiency in oral and written French by means of the MLA cooperative language proficiency test, French direct proficiency tests for teachers (TOP and TWO), and self-ratings. Unpublished Ph.D. dissertation, University of Minnesota.

Burger, S. 1989. Content-based ESL in a sheltered psychology course: input, output and outcomes. TESL Canada Journal/Revue TESL du Canada, 6:2, 45-59.

Clifford, Ray T. 1981. Convergent and discriminant validation of integrated and unitary language skills: the need for a research model. In Adrian S. Palmer, Peter J. M. Groot and George A. Trosper, eds. The Construct Validation of Tests of Communicative Competence. Washington, D. C.: Teachers of English to Speakers of Other Languages. Pp. 62-70.

Edwards, H., Wesche, M., Krashen, S., Clement, R., & Kruidenier, B., 1984. Second-language acquisition through subject-matter learning: a study of sheltered psychology classes at the University of Ottawa. The Canadian Modern Language Review, 41.2. 268-282.

Fouly, K., Bachman, L., and Cziko, G. 1990. The divisibility of language competence: a confirmatory approach. Language Learning, 40:1. 1-21.

Hauptman, P., Wesche, M., and Ready, D. 1988. Second-language acquisition through subject-matter learning: a follow-up study at the University of Ottawa. Language Learning, 38.3. 433-475.

Kramer, P. 1989. The Classroom Acquisition of German and the Input Hypothesis. Salt Lake City: University of Utah Ph.D Dissertation.

Kramer, P. and Palmer, A. 1990. Comparative program evaluation studies as tests of the Input Hypothesis. Paper presented at the TESOL '90 Convention, March 6-10. San Francisco.

Krashen, Stephen. 1985. The Input Hypothesis: Issues and Implications. London: Longman, Inc.

Lafayette, R. and Buscaglia, M. 1985. Students learn language via a civilization course--a comparison of second language classroom environments. Studies in Second Language Acquisition, 7.3.

Lightbown, P. 1989. Can they do it themselves? a comprehension-based ESL course for young children. To appear in the proceedings of the Conference on Comprehension Based Second Language Teaching: Current Trends. University of Ottawa, May, 1989.

Oller, J. 1979. Language Tests at School. London: Longman, Inc.

Oller, J. (ed.). 1983. A consensus for the eighties? In Issues in Language Testing research. Rowley, Mass.: Newbury House. 351-386.

Palmer, A. 1972. Testing communication. IRAL 10.1. 35-45.

Richards, J., & Rodgers, T. 1986. Approaches and Methods in Language Teaching: A Description and Analysis. London: Cambridge University Press.

Sternfeld, S. 1989. The University of Utah's immersion/multiliteracy program: an example of an area studies approach to the design of first-year college foreign language instruction. Foreign Language Annals, 22.4. 341-354.

Upshur, J., and Homburg, T. 1983. Some relations among language tests at successive ability levels. In Oller, J. (ed.). Issues in Language Testing Research. Rowley, Mass.: Newbury House. 188-202.

Upshur, J. and Palmer, A. 1974. Measures of accuracy, communicativity, and social judgements for two classes of foreign language speakers. Selected Papers from the Third International Congress of Applied Linguistics, vol. 2. Heidleberg: Julius Groos Verlag.

DEFINING AN APPROPRIATE ROLE FOR LANGUAGE TESTS IN INTENSIVE ENGLISH LANGUAGE PROGRAMS

Margaret Des Brisay
Doreen Ready

In the well-known play, "A Man for All Seasons", which tells the story of Sir Thomas More, there is a scene in which More advises a student of his to become a teacher. " You'd be a good teacher" he says to the young man. The young man replies "And if I were, who would know?"

"Who would know?" More specifically, "How would he know?" Researchers today collect a wide range of qualitative information to answer this question - teacher interviews, classroom observation, peer review, student course critiques - but when interest focuses on the effectiveness of the teaching in terms of measurable educational outcomes, then attention must be given to what the students can do as a consequence of the educational treatment they have received. Pretest-posttest measures of proficiency gains are frequently used to provide quantitative data for evaluating teaching effectiveness, not perhaps in the case of individual teachers but of groups of teachers and the programs with which they are associated. This is particularly true in the informal type of evaluation that goes on when funding agencies or their advisors select the in-country language school that will deliver what is known as pre-departure ESL training to their scholarship candidates. Although teachers and educational researchers are aware of the potential for the misuse of such data, to the lay person, gain scores seems the most the obvious way to determine the success of a teaching program.

In practice, even calculating gain scores seems unnecessarily complicated for many administrators. They prefer to look at pass rates which is understandable given that tests are used primarily to make decisions about individuals. There is some cause for concern, however, when language schools announce in their brochures that 78% of their students "successfully completed their course" which no mention being made of what criteria were used for measuring success and when these were met on schedule or not. It could simply mean that 78% lived to tell the tale...eventually. One formal evaluation of a program with which we were associated featured a chart showing the percentage of students meeting the exit standards in several consecutive semesters. This chart was used to compare the performance of different directors although it made no mention of changes in the length of the semesters, a steady rise in the entry level of the students and compensatory adjustments made to exit scores. If pass rates are to be used to compare programs, they must be interpreted with reference to entry level and other baseline data which, in effect, leads you back to gain scores.

While we concede that test scores have a legitimate role to play in evaluating teaching effectiveness, what follows is , in fact, a cautionary tale, the moral of which is that test scores are not as simple, clear and conclusive as advocates might wish to believe.

CONTEXT FOR STUDY

For the past two years, researchers at the University of Ottawa have been conducting analyses of ESL test scores obtained by students prior to and during their training in several different intensive English programs in Jakarta, Indonesia. The examinees in all cases were candidates for scholarships to either the United States or Canada who had satisfied all the selection criteria for such an assignment except for the English language requirement. This language requirement was defined as achieving a level of English proficiency adequate for academic purposes and was re-stated in terms of a test score of 550 on an International TOEFL.

In other words, not only was the TOEFL being used to measure attainment of the instructional objectives, in many ways, the TOEFL was the instructional objective.

The initial Canadian ESL program in Indonesia had produced disappointing results with many fewer candidates than anticipated reaching the predicted TOEFL score within the allotted time and some being dropped from the program entirely. Moreover, the broadly-defined objectives of the program, preparing students for study abroad, frequently conflicted with the narrowly-defined objective of getting 550 on the TOEFL. However, as one ministry official said when it was suggested that the future needs in an academic environment should be given priority in the program, "Before they can succeed, they must be admitted, and before they can be admitted, they must have 550 on the TOEFL."

The study to be reported in this paper was one of several undertaken in order to try to address the concerns of Canadian program planners about the Indonesian training model. Specifically planners were interested in knowing whether 1) realistic estimates of training requirements were being made 2) whether data from other tests, in particular, the Canadian T^st of English for Scholars and Trainees, would enable better predictions about end-of-course success to be made, (and hence, better initial selection) and 3) whether some guidelines could be established for striking a balance between test preparation and preparation for life after the test.

(The comparability study between the two tests has been reported on elsewhere and will be referred to only briefly in the present paper. And it should be stressed that there was never any intention of using the results of this study to evaluate the centres or teachers concerned but rather to use then in setting reasonable objectives for future programs. What sort of gains is it reasonable to expect? Who is most likely to succeed in the time allotted? How many are likely to succeed?)

METHODOLOGY

There were 129 subjects in the study , spread over 8 classes (average of 16 students per class) in three different language centres. Two classes were at Centre A, three at Centre B and three at Centre C. All subjects were studying at the EAPII level. A wide range of test data was collected but the discussion will focus on the test scores from two tests, an International TOEFL written in October 1988 and a second International TOEFL written in January 1989. The October tests were written after students had been studying in a TOEFL preparation program for 12 weeks and January TOEFL following an additional 11-12 weeks of instruction that emphasized academic skills.

Subjects could not be randomly distributed among the three language centres. It appears, however, that students came from similar cultural, linguistic and educational backgrounds. Twenty-six had been studying in intensive ESL programs since January 88, fifty-seven had begun studying at the EAPI level in April and forty-six had tested in when the EAPII program began in July. (No information was available about the previous ESL training of this latter group.) TOEFL entry scores indicated that the range of scores was similar at each of the three language centres although subsequent analysis showed this apparent homogeneity to be a bit misleading. (See discussion below.)

INSTRUCTIONAL TREATMENT

The instruction at each centre was guided by the same general objectives (academic readiness and TOEFL preparation) but , in principle, no single methodology was imposed. An examination of the end-of-course reports does suggest a differential emphasis on TOEFL preparation. One centre (Centre A) administered scores from 22 institutional TOEFL's with each TOEFL followed by a thorough post mortem, implying a minimum of 20% of classtime was devoted to actual TOEFL practice. The fluctuating nature of the scores must have brought tears of frustration to all concerned although they would not have impressed anyone cognizant of the standard error of measurement and the fact that measurable gains in overall proficiency are not registered weekly. Centre B reported administering 13 practice TOEFL's, mostly in the weeks prior to the October TOEFL while a third (Centre C) reported only three although students were known to have done a lot of practice tests on their own. Two centres give detailed information and assessments from a writing course which took up 25% of classtime. All in all, enough differences were reported in course content that it was decided to further analyze the data to investigate the impact, if any, of these differences on the test scores.

AN EXAMINATION OF REPORTED SCORES

TABLE ONE: Overall Results on TOEFL Total and Part Scores: (n=115)

	JULY	OCT.	JAN.
Total	499.0	527.0	533.6
Score	(18.0)	(32.6)	(27.0)
Part One	47.9	52.5	52.6
Listening	(3.1)	(4.6)	(2.8)
Part Two	50.3	53.5	53.6
Structures	(3.5)	(4.7)	(3.6)
Part Three	51.3	53.4	54.7
Vocal/Read	(3.3)	(3.6)	(3.6)

(-) = sd

Table One shows means and standard deviations for totals and part scores on the three TOEFL's for the 115 students for whom there were no missing data. The standard deviations for the group show that the relative homogeneity suggested by the July scores is not maintained in October nor in January. In fact, the standard deviation (an indication of the range of scores) almost doubles. It should be noted that the "July" TOEFL scores were obtained on institutional TOEFL's that had been written at different times.

It can be seen that the mean gain on total scores was 34 points, this being composed of a gain of 28 points during the first 12 weeks which were largely devoted to TOEFL practice at all three centres delivering instruction and 6 points during the remaining 12 weeks where the emphasis shifted to academic preparation. Thirty percent of the students actually achieved the exit standard of 550 on the October TOEFL while 31 percent did so in January.

The standard deviations of 32.6 and 27.0 for the two International TOEFL's in **Table One** indicate a great deal of individual variation. Individual changes ranged from a gain of 47 points to a loss of 50. In fact, 37 students actually had lower scores in January.

FURTHER ANALYSES

A regression analysis was performed using the Part TOEFL scores obtained in October and the part CanTEST scores obtained at the same time in order to try to arrive at the best equation for predicting the TOEFL scores obtained in January. The analysis excluded the 39 subjects who had obtained 550 on the TOEFL of October since it was felt that they might not be as motivated as the others.

The data were further analyzed using multivariate analysis of variance in order to see if there were any statistical differences among the three centres delivering instruction with regard to the part TOEFL scores on January 22nd. (This procedures also took into account differences in Part TOEFL scores obtained on October 14).

The analysis was repeated using the eight classes as the independent variables instead of Centres to see if classroom variables might have affected gain scores.

REGRESSION ANALYSIS

The results of the regression analysis in which the part CanTEST scores and the part TOEFL scores (October88) including the Test of Written English (TWE) score were used to try to predict the January TOEFL score are reported in Table Two. The method used was stepwise regression in which variables are entered into the equation until there is no further increase in multiple R.

Although this and similar studies (Des Brisay, 1989) show reading scores to account for more of the shared variance than any of the other predictors, their relationship to the dependent variable is not strong enough for them to be used with any confidence to predict success. Moreover, the variables entered into the equation only account for about half the variance present. The total variance present is a measure of how much individual scores vary from the group average. This means that although the prediction equation in **Table Two** gives some idea of what the final TOEFL scores will be there is still a large amount of error so that ESL program planners cannot count on scores obtained before training to give a really accurate prediction of the outcome of training.

TABLE TWO: Regression Analysis to Predict Final TOEFL Score (January) from previous Part TOEFL Scores and Part CanTest Scores(October).

DEPENDENT VARIABLE: TOEFL TOTAL (JAN 89) (N=129)

Predictors	b
TOEFL Reading	3.137
TOEFL Structures	1.806
TOEFL Listening	1.446
Constant	193.887
Multiplier	.717
R^2	.514
Standard Error	16.58

(These results can be contrasted with those obtained when a similar analysis was done in another program (Des Brisay, 1989) where CanTEST was being used in the decision making and incoming students had had limited experience with the TOEFL. In this case, it was CanTEST reading scores which were the best predictors of final TOEFL scores.)

GROUPS FORMED ON DATE OF ENTRY AND ENTRY SCORES

In order to see whether it might be possible to control some of the sources of individual variation, the data were examined to see if differences in either gain scores and/or pass rates could be related to individual differences in proficiency at entry or the length of intensive training as measured by date of entry into the program. Descriptive statistics for groups formed by date of entry and by TOEFL1 (July) scores are shown in **Tables Three and Four.**

In **Table Three**, we see that students who tested directly into EAPII(Group 3) had higher means and more successes than those who were promoted in from EAPI (Group 2) while these in turn had better test performances than those who had previously done both BELT and EAPI (Group 1). (Only the differences between this latter group and the direct entrants were significant and then only for the listening and reading sections.)

Table Three: Means and Standard Deviations on Part TOEFL Scores for Groups Formed on Date of Entry.

	October	January	Gain
Group One: 01/88	515.64	527.05	11.4
(n=29)	(33.8)	(27.6)	
Group Two: 04/88	521.75	530.9	9.3
(n=57)	(32.2)	(27.9)	
Group Three: 07/88	539.8	540.51	.71
(n=52)	(28.7)	(26.7)	

Table Four: Means and Standard Deviations on Part TOEFL Scores for Groups Formed on July TOEFL Scores.

	October	January
Group One:(below 500)	520.80	530.05
(n=69)	(29.36)	(28.11)
Group Two: (501-525)	541.6	542.79
(n=51)	(27.4)	(24.52)
Group Three: (over 525)	549.33	548.66
(n=14)	(23.4)	(23.05)

However, when we look at the gain scores by date of entry we see that the students who began their intensive instruction in January are making larger gains even though they are still farther below the exit standard; they are making larger gains as a group partly just because they are weaker and students in the lower score ranges typically register larger gains.This difference by level reflects the fact that test scores are not truly equal interval in terms of knowledge increment. An comparison of the gains made by three groups formed on the basis of their initial TOEFL scores (Table Four) supports this in that larger gains are observed among the lower proficiency groups.

There would be no way to further explore sources of individual differences without more knowledge of the previous language learning experiences, general intelligence and particular learning styles of this group of students. (Scores from an academic proficiency test were available and correlated at .07 with TOEFL entry scores and .33 with January TOEFL scores).

CENTRES AND CLASSES AS VARIABLES

Table Five: Means and Standard Deviations for Part TOEFL Scores October and January by Centre.

	Listening		Structures		Reading	
CENTRE	Oct	Jan	Oct	Jan	Oct	Jan
A (n=31)	52.4	53.2	54.3	54.7	54.1	55.8
	(4.1)	(3.5)	(4.4)	(3.5)	(3.8)	(3.6)
B (n=50)	51.2	51.8	52.4	52.9	51.3	53.3
	(5.0)	(4.4)	(5.5)	(3.3)	(3.7)	(3.5)
C (n=48)	53.2	52.6	52.3	53.1	54.0	54.4
	(4.4)	(3.0)	(4.4)	(3.8)	(3.2)	3.5)

The results of the multivariate analysis do not support any conclusions about the efficacy of any one Centre over another [illegible]e analysis does support the conclusion that Centre A is stronger but this is true in October as well as January. There were no significant gains made by any Centre on the listening and structure section of the test and all three made gains in reading which did not differ significantly from each other.

The matter of statistical significance is very important considering the decisions that may be made on the strength of the appearance of differences. It should be kept in mind that group average scores, such as those in Table Five, are made up of individuals scores which may vary considerably from the average score. These individual scores , depending on whether they are well above or below that average score, can raise or lower it accordingly. Thus, although there may appear to be between group differences , once the group mean scores have been analyzed using rigourous statistical methods, these differences become something attributable to chance alone; in other words, the differences are not statistically significant.

This lack of statistical significance is not entirely unexpected especially in view of the fact that students are never randomly distributed to training groups and there is no guarantee that the groups being compared were ever equal before training began.

Table Six: Means and Standard Deviations for Total TOEFL Scores October and January by Class

CLASS	JULY	OCT.	JAN.	Raw Gain	Pass (%)
1 (16)	491.0 (11.6)	532.0 (27.6)	543 (23.7)	11	50
2 (17)	489.0 (12.9)	540.0 (29.8)	546.0 (32.4)	6	44
3 (18)	503.0 (19.3)	507.0 (36.3)	523.0 (30.5)	16	17
4 (16)	495.0 (20.6)	509.0 (31.8)	519.0 (23.7)	10	12
5 (14)	511.0 (17.2)	537.0 (34.9)	538.0 (26.8)	1	50
6 (16)	496.0 (17.3)	519.0 (36.4)	533.0 (26.3)	14	44
7 (17)	504.0 (19.8)	542.0 (22.9)	538.0 (23.7)	-4	13
8 (18)	501.0 (17.8)	530.0 (26.2)	529.0 (28.4)	-1	23

Table Six gives similar statistics for all eight classes. As previously noted, the scores for TOEFL1 were obtained on institutional TOEFL's and were not all written at the same time. However. all students wrote a version of the Canadian Test of English for Scholars and Trainees within the first week of their program and the classes are ranked in a similar way according to the CanTEST results. As was the case with Centres, none of these differences is statistically significant.

Some researchers would question whether raw gains should be used at all to measure growth in instructional settings, much less to make comparisons among different groups since raw gains typically level off as students become more proficient. Swinton (1983) describes one possible source of error in calculating gain scores when there is a wide range in scores. That is the statistical phenomenon of the regression to the mean. With this data, that is not a threat because the range of scores is extremely narrow (475-525)

IMPLICATIONS OF THE REPORTED SCORES FOR THE TRAINING MODEL

As stated initially in this paper, the data were not collected for the purpose of evaluating the teaching at the different centres but rather for evaluating the training model itself. In this context, the study clearly shows the need for better baseline data. The means and standard deviations for the entry TOEFL (**Table One**) give a misleading impression of the homogeneity of the group, even allowing for the imperfect way this can be reflected in any test score. Although it is commonly found that students will progress at different rates so that the range of abilities in a given group may increase over a period of instructions, nevertheless, the July scores, which were obtained on a number of different institutional TOEFL's written at different times, do not provide adequate baseline data for determining progress. This is a finding that can be easily operationalized. However, a more controlled selection process should not be undertaken for the purpose of keeping people out but for making more realistic estimates of training requirements.

Improving the pass rate within the present time frames would involve insisting on higher entry scores. This too could be easily operationalized but would seriously reduce the pool of potential candidates and risk putting concerns about costs of language training ahead of the larger aims of such technical assistance programs which imply giving an equal opportunity to all otherwise qualified candidates. Moreover, although the perception of the teachers that continuing students are somehow weaker is supported by the findings , this can in no way be interpreted to mean that as a group they are poorer language learners. Their poorer performance simply reflects the fact that as a group, they were only minimally proficient for EAPII on entry and had further to go to reach the exit standard.

The test data do not permit any useful comparisons to be made among the centres involved. The observed score patterns might well be interpreted differently if less refined statistics, such as pass rates or gains on total scores (enlarged by the ETS practice of multiplying everything by ten thirds) were used. In that case, some classes and some centres could appear to have been more successful than others. The percentage of students achieving the desired TOEFL score did vary from class to class (50% to 14%) and centre to centre (45% to 27%) but as we have noted above, following multivariate analysis, none of the gains on part scores shown in **Tables Five and Six** were found to be statistically different by class or centre. The difference in pass rates, then, could be equally well attributed to chance and/or to the characteristics of the class on entry. The extent to which administrators would be impressed or distressed by the score patterns revealed in this study would partly depend partly on their degree of statistical sophistication.

IMPLICATIONS FOR THE USE OF GAIN SCORES IN PROGRAM EVALUATION

In the particular program under study, efforts had been made to avoid the methodological weaknesses that have plagued other attempts to quantify teaching effectiveness. Classes were of similar size with a similar balance of continuing and newly placed students. Instruction was of the same length and intensity, and as previously mentioned, students were of similar educational, cultural and linguistic backgrounds and students were thought to be at similar levels of proficiency on entry.

It is individual differences in proficiency gains as measured by the TOEFL which are the dominant finding of this study. Whatever group tendencies can be found are of limited use in program planning and of virtually no use in program evaluation. We can estimate from this and other similar studies that approximately 1/3 of a group of students studying at the EAP II level will reach TOEFL 550 after 18 to 20 weeks of intensive ESL instruction but which ones they will be cannot be predicted from the test scores. (Probably the teachers know, but how do they know?)

The fact that no statistically significant differences among centres or classes were found may simply suggest that all centres were equally effective (or perhaps, from a sponsor's point of view, equally ineffective). However, the fact that group means disguise so much individual variation and the testing instrument used failed to measure the learning that must have been taking place does have implications for future efforts to use gain scores as a measure of program effectiveness. Such efforts will have to recognize, as educators have always done, that:

> no treatment can be equally appropriate for everyone and as a corollary to this, similar instructional treatments will have a wide range of outcomes. We may be able to say that 35 to 40% of students entering an intensive ESL program at the EAPII level will reach the exit standard of 550 on the TOEFL with 22 weeks of instruction, but which ones they will be, we cannot say;
>
> general proficiency tests are not appropriate for measuring gains over short periods of time (if 360 hours of intensive training can be considered short) and, moreover, such tests will be particularly insensitive to growth in specific skill areas such as writing for academic purposes;

While educational researchers consistently stress that evaluation cannot be based solely on testing student product, (Weir, 1989), program accountability does seem to require that an instructional program have measurable educational outcomes . Given that it would not be cost-effective to provide individualized instruction and assessment, then clearly more appropriate testing instruments an statistical techniques are required.

Bachman (1986) optimistically declares "new developments in criterion-referenced test theory and more comprehensive definitions of language proficiency provide keys to developing language tests that are appropriate to the needs of language program evaluation." Developing such a testing system takes time and a good deal of money. Even when the reliability and validity of the new instrument has been empirically established, one must still establish its credibility in the eyes of the gate-keepers to North American universities. It becomes a question not of "Who would know?" but "Who would believe you?"

The poor performance of these 129 subjects on the January TOEFL offers a compelling argument against the use of a norm-referenced standardized general proficiency test to measure achievement in an academic skills program. You will recall that there was group gain of only 6.2 points and perhaps the most striking finding in the study is the large number (48 out 129) of students who actually had lower TOEFL scores in January than they had had in October, something that cannot satisfactorily be explained away by referring to standard error and regression to the mean.

It is not unreasonable to assume that the 39 students who had achieved their exit score in October were less motivated to do well on the January TOEFL. Here, as elsewhere, there was great individual variation. Twenty-one of the students scoring 550 or more in October had lower scores on the January TOEFL, 3 remained the same, 12 improved and three others did not (wisely, perhaps) write the second TOEFL. On the other hand, 24 students who had not passed in October also had lower scores in January, a phenomenon not likely to be explained by a decrease in motivation.

Twenty two of these "losers" were students in Program C, the least TOEFL intensive of the three centres. Neither the possibility that the January test was easier or that nothing was learned can be seriously entertained. The fact that the "losers" tend to be concentrated in the centre providing the least TOEFL practice between the two tests and the "winners" in the program providing the most, suggests an attractive line of inquiry that would be impossible to pursue on the basis of the data available. It is tempting to suggest group differences might have been more marked had not the need for achieving a certain TOEFL score been uppermost in the students' minds . Given this pressure to pass the TOEFL, they may have simply selected from the different programs whatever they thought would be useful to them in achieving this end and did not fully engage in the rest.

Even in studies where differences in gains and successes can be shown to be statistically significant, it is difficult to trace causal relationships. To quote Long (1983), "We often don't know if he gained because of the program, in spite of tne program or merely while registered in the program." When it comes to choosing an institution to deliver ESL instruction any informed program planner knows that other information must be collected. Canadian planners, for example, observe classes, examine curriculum documents, evaluate façilities, such as a libraries, provisions for self-study, support staff, language labs, look for resources that will provide cultural and academic preparation in addition to language training and do not allow themselves to be unduly impressed by claims of a high pass rate or promises of dramatic gains. It is not unknown, however, for groups of students to be moved from one centre to another or for individual students to be dropped from a program because improper inferences have been drawn from test results.

CANADIAN INITIATIVES IN ESL TRAINING AND TESTING

The Canadian International Development Agency funds several Human Resources Development Projects that have a language training component. The goal of the latter is to select, train and certify candidates from the developing countries who wish to come to Canada for either university study or practical attachments. As with many similar development programs, planners have had to face the fact that the greater the number of stakeholders, the greater the need for some form of standardized evaluation to provide for comparability among programs and overall program accountability.

Fortunately, they have also come to appreciate the extent to which a certification test can "steer the curriculum" (Canale 1988). In order to ensure positive washback from the test used, CIDA has provided financial support for the development of a program-specific testing system. This is clearly not a solution for everyone. I mentioned some of the problems above. However, CIDA is acutely aware of how inaccurate assessments of ESL proficiency can result in unexpected expenses to the funding agency as well as lost opportunities for otherwise qualified scholars and trainees.

The Canadian Test of English for Scholars and Trainees is compiled from an item bank consisting of authentic texts for both listening and reading comprehension and is supplemented by a writing exam and an oral interview. The fact that more information about language proficiency is available when making the initial selection and that students must be at least at a level corresponding to EAPI means that nearly all non-academic track candidates (trainees) are able to reach their exit standards in an 18 week semester while academic track candidates (scholars) generally require two semesters.

Test reaction questionnaires are completed by both teachers and test takers following each administration. This input, plus continuing dialogue with teaching staff and materials developers help strengthen the alignment between curriculum and tests so the tests can more credibly measure attainment of the instructional objectives.(Gatbonton, 1989, Des Brisay 1989) For example, there are no single sentence prompts, no isolated grammar or vocabulary questions on the CanTEST as teachers found this discouraged students from dealing with longer contextualized samples of language. Finally, the fact that the tests are normed on specific sub-sets of the international test clientele permits a more sensitive interpretation of scores.

We would like to mention briefly three other programs which provide alternative models designed to lessen overdependence on test scores for making consequential decisions. In one program , students are relieved of the necessity of writing any ESL tests beyond the first one. A thorough diagnosis is made at entry and generous training estimates are made. (After all, you can always send someone abroad earlier than anticipated but it is demoralizing to keep him or her back.) No formal testing is done again after the initial projections so that the instruction can focus on preparing students for the fùture. Although the CanTEST is administered to provide for program accountability (and comparability), decisions affecting the students are not based on CanTEST scores, more or less eliminating test anxiety. Such a program is only possible because a small group (15 per year) of students is involved, special arrangements have been made with the admissions office at their Canadian university and administrators are prepared to offer any necessary post-admission ESL support.

Another program recognizes the fact that language proficiency takes a long time to develop and it may not be cost-effective to keep a student in language training until he has reached a proficiency level adequate for the writing of a Ph.D thesis. In this program students begin their course work in their own country with visiting Canadian professors before coming to Canada for 12 months of study. They then return home to write their theses in their mother tongue. The degrees are joint degrees (Ph.D in management) granted by the Canadian and Chinese universities involved.

Yet another program which allows for the steady but slow maturation of ESL proficiency without excessively delaying academic training involves a teacher training program for future ESL teachers in Malaysian secondary schools. By selecting recent high school graduates for this program, the high cost of removing an active professional from the work force for lengthy periods of language training is avoided. The students do all their undergraduate training in Canada but Canadian faculty are counselled on how to evaluate their work in spite of ESL problems and marks assigned in the first two years make allowances for communication problems related to ESL proficiency.

And finally, recognizing that the information requirements of sponsors and admissions officers will dictate the continued use of standardized tests for certification in most programs, TESL Canada is trying to encourage the informed use of a wider range of ESL admissions tests in Canadian post-secondary institutions through the production of a manual for test score users. The proposed user's guide will explain how different tests relate to each other and contain details concerning the reliability, accessibility, quality, significance and security of the information of each test. The TOEFL, the CanTEST, the new IELTS, the Ontario Test of English as a Second Language (OTESL), the University of Toronto's Certificate of Proficiency in English (COPE) are among the tests to be included. It is hoped that this manual will enable programs which do not have the resources to develop their own test to at least pick the one closest to their needs with confidence that the scores will be recognized by receiving institutions.

In closing let me finish the story of Sir Thomas More and his student. When the student complained that no one would know if he were a good teacher or not, More replied. " You will know, and your students, and God. That's not a bad audience." Unfortunately, the audience does not seem to have enlarged much since More's time and since God is not available to work on evaluation teams, we must look to other authorities to satisfy the information requirements of external stakeholders. This demands new models for evaluating instructional programs in which the role played by test scores must continue to be interpreted carefully and cautiously.

REFERENCES

Bachman, L.B. (1989); The Role of Criterion-referenced Test in Language Program Evaluation,in Johnson,K., eds. .....Cambridge University Press

Canale, M. (1987). Language assessment: The method is the message. In D. Tannen and J.E. Alatis (eds), The interdependence of theory, data, and application, Washington, DC: Georgetown University Press. 249-262. (Georgetown University Round Table).

Canale, M. (1988). The Measurement of Communicative Competence, in Annual Review of Applied Linguistics, 8, 67-84. Cambridge University Press.

Des Brisay, M. (1989). The Problem of the Middle Ground: where do you draw the line? Paper presented at the 11th Annual Language Testing Research Colloquium, San Antonio, March 1989.

Gatbonton, Elizabeth (1989). Taxonomy of Language Tasks and Objectives: English Course for Chinese Visiting Scholars, Trainees and Consultants in Canada. Document prepared for the Canada China Language and Cultural Program, St. Mary's University, Halifax.

Long, Michael,(1983). Presentation at TESOL Summer Institute , University of Toronto , Toronto, July 1983.

Overseas Training Office, Government of Indonesia(1988). Preparing Indonesians for Overseas Training. Presentation at the National Association for Foreign Students Affairs Conference, Washington, DC. June 1988.

Swinton,D. (1983). Measuring Growth in Instructional Settings.TOEFL Research Report No.13.Educational Testing Service, Princeton,

Weir, C. (1989); Program Accountability;The writing is on the wall; unpublished document, Reading University .

PRINCIPLES AND PRACTICE IN AN EVALUATION PROJECT

Dermot F Murphy

INTRODUCTION

This rather grand title masks a basic problem: how do we put our ideas into practice and get them to work? More importantly how do we get other people to put them into use and make them effective? I want to look at these questions in the context of an evaluation project I have been contributing to on behalf of the Brtitish Council and Overseas Development Agency. The project was set up in the Schools Division of the Malaysian Ministry of Education and began work in January 1989; it is only in its initial stages. None of these bodies is responsible for the opinions that I express here, though I hope that they might agree with most of them!

The Project was motivated by the introduction of the Kurikulum Bersepadu Sekolah Menegah (KBSM), the five year Integrated Secondary School Curriculum announced by the Malaysian Ministry of Education in 1987. The curriculum was prepared by the Curriculum Development Centre, which is in charge of policy, new curricula, as well as their introduction, and through the State Education Offices, of the INSET to support their introduction. The curriculum contains a statement of aims and content. It is being introduced a year at a time, and started with language course in January 1988, so at the time of writing, the third year curriculum for English is being used for the first time. The implementation of the curriculum, and the administration and management of schools is the responsibility of Bahagian Sekolah-Sekolah, Schools Division, which works through the various State Education Offices. A committee consisting of officers from several Divisions of the Ministry is responsible for producing a handbook for teachers on the methodology to be used (Goh et al 1989).

In establishing this project the Ministry had a number of aims and procedures in view. It wants to be able to assess the effects of the new curriculum on teaching and learning in the classroom; it wants to be able to gather contributions from teachers themselves to further development of the curriculum, since, it is hoped, evaluation will provide more accurate information on learners' needs, among other things. The Ministry also wants to involve teachers in the process of curriculum development, and in addition to ensure that appropriate in-service education is provided. In order to achieve these aims it decided to concentrate on formative (ongoing) rather than summative (end of course) evaluation, focussing more on the processes of the implementation than on the final product of the curriculum. Whether the evaluation is formative or summative is decided by the evaluators' aims and the use made of findings more than by other factors. The project chose this focus on formative evaluation because it was felt that summative procedures would deliver some of the information too late and in a form where it would be difficult to account for how the teaching and learning proceeded.

My contribution to the project has been to conduct two one-week seminars at eight months' interval, followed by short periods of field work. The participants in the seminars came from the different state education offices and from the Ministry. They included State Language Officers, Supervisors for English. Resource Personnel and teachers. This paper describes my contribution. I will outline some of the ideas that I feel are guiding the project in its first stages, and describe my input and findings. In essence this is an essay about change, a case study on the beginnings of one innovation.

Background

Evaluation is the process of assessing what you are doing to see how worthwhile it is; the action may be assessed in terms of cost-effectiveness, of attainment matched to normative goals, or it may be done in a goal-free approach seeing whether what is being done has value, particularly in the participants' view, from an ethnographic standpoint. At times evaluation will be called action research, it is about applying research techniques to find out things you need or want to know. It may be done as part of a national scheme, or by one teacher with one class. Many issues surround evaluation: the reasons for doing it, its timing and duration, its scope, the methods to be used, who to involve, and these points need to be considered here.

One problem with evaluation is that it seems to raise more questions than it answers, and even then the answers may raise further questions. So is it worthwhile? Do we need it? There is a lot of evidence to suggest that attempts to improve education generally have little success (Holt 1987); centrally planned change rarely produces the desired or expected results. We learn this from evaluation, usually summative, even terminal. I have suggested elsewhere (Murphy 1985) that sometimes changes in ELT have not delivered expected innovatory results because they were introduced on the basis of plausible but speculative proposals. Usually these proposals did not mention how they might be evaluated, and those who were implementing the ideas did not include an evaluation scheme as part of their curriculum design.

Another problem with following the latest speculative change is that however principled its theory, a logical argument is no guarantee of operational success. ELT has changed its approach as if following intellectual fashion, just as from time to time charismatic movements have had widespread influence (cf Murphy 1981). Often the theoretical proposals of these movements contain good practical sense, based on sound technological experience, and on balance I feel that we have been making progress, clarifying our ideas about what it is we are trying to do.

However, there is also a danger that we have missed valuable insights, and abandoned practice without properly assessing its worth. So swings to the latest approach have contributed to a process of change and development by revolution: this year's innovation rejects last year's doctrine. Is this an effective way to produce change? It does not seem to have been so if we judge by the continued dissatisfaction with results in language teaching (it is not simply confined to ELT). Where are its weakness? They spring from the proposal being unquestioned rather than experimental. Though the lack of experiment may have as much to do with the implementer as with the proposer, it should be said.

What is forgotten is that the approach needs to be adapted to the local circumstances and context of its use. These factors will influence what is taken as the scope of evaluation, because another problem is that we could evaluate every aspect of the implementation. Which would probably bring all the work it was focussed on to a halt. Issues and areas must be identified, after which proper sampling techniques and distribution of work and responsibilities will allow broad scope. Evaluation is done to avoid being wise after the event; hindsight is a powerful analytical technique, but its findings usually come too late, when we are disillusioned with last year's grand proposal. So when is it best to evaluate? It can be done from the beginning of an innovation.

What should change be like then? I suggest that evolutionary change is more likely to succeed, but what does this metaphor mean? It would require assessing the worth of what we have and already do before deciding to add new practice and see how the innovation works. Why is this not just another speculative proposal? It is not speculative because it says that it should be tried out and measured. What does it imply for those in and trying to achieve change? Basically, the idea that change needs to be managed and evaluated. Over recent years this notion has gained considerable currency in ELT, catching up with practice elsewhere in education (eg Alderson (ed) 1985, Nunan 1988, White 1988). Then how is the effectiveness of what we do and of the change to be measured? Who guarantees the measure? These questions are not so easy to answer, but they must be faced.

A different problem arises in that evaluation undertaken on this scale is so frequently seen and done as a project alongside the curriculum rather than as part of it; evaluation has become its own separate discipline, outside the mainstream. Should it be a separate enterprise from the rest of the curriculum? This suggests that it requires expertise to operate evaluation. If it is not to be left as the domain of a few experts, how is it to proceed? Is it safe to let it into the hands of what some would deem semi-skilled users? As you can see, there are several questions here already; essentially they are concerned with why and when we should be doing evaluation, how, and who should be doing it. I now want to describe how I have been answering these questions for the project I am concerned with. My remarks will be grouped under points about innovation, about the underlying principles and about practice.

Innovation

The KBSM is innovatory, so we have had to discuss the nature of innovation, and in the context of this project, the role of evaluation in promoting it. Change may occur as part of the passing of time; there is no conscious attempt to influence activity in particular ways. In talking of innovation we refer to change which is planned; an innovation is a deliberate attempt to alter materials or practice in one or more ways. In this case you need to know what is going on, and obtain information to show the effects of the innovation and if necessary to serve as a basis for adjusting or modifying the planned action. Note that you do not have to be carrying through some innovation to do evaluation, though people seem to think of evaluating more often in association with new practice. As I said earlier, when the decision to evaluate is an afterthought it may come too late to do more than note that the innovation did not succeed, so evaluation needs to be part of the innovation from the first if it is to fulfill its role of monitoring and informing. This implies that innovation has to be managed, and that evaluation can supply the information necessary for the management process.

There are implications for the management style, and for what is done in evaluation, depending on the origin of the innovation: whether it is top-down or bottom-up. Much of the innovation in education is top-down: it comes from Ministries or Development Centres, plans being handed down for implementation. Examples of bottom-up innovation such as the graded-objectives t·sting movement in foreign language teaching in Britain, or the original RSA Dip TEFL, a teacher qualification proposed by a London college for validation by the Royal Society of Arts Examination Board are rare. Polar models such as the top-down bottom-up metaphor suggest two opposing approaches, whereas in reality we find that the source of action and certainly its focus are more accurately located on a cline.

Nevertheless, there is a widespread perception that change initiated from below is more successful than change initiated form above; this oversimplifies the process. The important element is that the focus, the activity of innovation is at the bottom, in the grassroots, even if the initiative came from above. When change does come from below, it eventually needs to be accepted and taken up by those above in order to ensure adequate support for development and diffusion of the innovation. In onecurriculum project, the link between those working below was not made with those above, with the consequence that the team was later deemed (by the top) to be "out of touch with ordinary teachers" - exactly the sort of people who made up the team. Consequently the work of the project did not get disseminated.

Can we explain why bottom-up innovation is perceived to be more successful in achieving results? The impact of change is noted at all levels in all spheres of life: the disruption of change and resentment of its effects are reported from many sources. It seems that we do not like change that is imposed on us and for which we can see no value. People usually have a number of question about innovation: who is promoting it? What is in it for me? What can you tell me about it? The points at issue then, are attitude to the innovation, ownership of the innovation, its value, and communication about the innovation.

In bottom-up innovation in education some of the people most affected are involved in creating and promoting the innovation; these are the teachers. They own the innovation when, for example, they are involved in writing and piloting new teaching materials. The value of the innovation is immediate because they are doing something which they perceive as being adapted to their professional needs. So if they want teaching materials which are better suited to their pupils, and to the curriculum aims, materials which are more lively and stimulating, and they are creating them themselves, then there is a tangible return. Often this will take the form of enhancing or upgrading their professional skills, a return which has considerable personal value.

Contact with the innovation will form teachers' attitude towards it and its effects, and their attitude is more likely to be a positive one if they feel they have some control over what is done. Some teachers and outsiders will have a negative attitude towards the innovation, criticising it for sound or personal reasons. The teachers may not like materials which expect them to master new management techniques, or which do not contain the subject content they believe is appropriate. In bottom-up innovation they are surrounded by others who can communicate their views, so there can be a real debate over what is being done. In top-down change their dissatisfied views may be more readily listened to and even become a leading influence.

Communication about innovation needs to be general; it is not enough for those immediately involved to keep in touch. They need to be informing who might have an interest in what they are doing: sponsors, colleagues, associated departments, parents, pupils. The innovators should tell them why the innovation has been introduced and what the benefits are. When this is not done you get the kind of result I mentioned earlier; resentment of an exclusive, secretive group, which may lead to its work foundering. The top is likely to resent bottom-up innovation just as the bottom resents top-down innovation.

Another reason for bottom-up change being more likely to succeed is that any innovation carries a cost. At the implementation stage teachers will have to be prepared to attend information meetings, and in-service training sessions; their workload may increase as they have to find new materials, or complete new administrative procedures. The introduction of the new National Curriculum in Britain is demonstrating all these effects. The cost seems less if you are benefiting and you are creating it because you are responsible for the innovation. The enthusiasm of a group of involved people will carry a great burden even over several years as I saw in one project.

The kind if innovation that is prepared at the top by a specialist group may be removed from the reality of many individual teachers' classrooms. The ideal plan in theoretical terms may not be suitable for a deprived urban school where the children do not speak the national language and have ambivalent, even hostile attitudes towards education in any case. Schemes for innovation will be modified in practice, or ignored if they appear incapable of adaption; consequently schemes need to be designed to allow for modification and reinterpretation.

This suggested capacity for adaption will only be made effective through evaluation. A fixed, monumental curriculum does not include evaluation: its ethos is against it. Always top-down, such curricula are authoritative and normative; they may be ignored by teachers or serve as a source of anxiety. They are inefficient and ineffectual: more than one teacher in these circumstances has said to me, "We are trying to finish the curriculum rather than teach the learners".

A curriculum which is a working, evolving plan needs formative evaluation to provide the information for modification and development. The findings of evaluation may include surveys of attitude, particularly where they may reveal problems, or on the other hand progress in getting the change accepted and adopted. Developing the ownership of those involved will be done through getting them to evaluate the materials they are producing and piloting, as well as their developing mastery of new skills, such as using unfamiliar teaching techniques. Finally, much of the information for communication about the innovation will come from evaluation findings. This discussion has set out the role of evaluation in innovator change; it has gone part of the way towards answering some of the questions raised initially, though we still have to show they fit with the formative evaluation of English on the KBSM.

Principles

Opening the second training seminar for the project, the then Head of the Language Unit in Schools Division said that he hoped that formative evaluation would become a standard part of practice for teachers of English in Malaysian schools. This long term aim for the project sets a direction for the principles which guide its establishment. Let us turn again to those initial questions.

Why do evaluation if it represents a cost as described above? The answer to this is short: the cost of doing evaluation is less than the possible cost of getting the overall project wrong and of coming to feel that you need to start all over again. However, there is a more assured return also: that if you are doing evaluation then you will have greater control over the implementation of the curriculum. It will create more accountability: make the implementers at all levels see the events of the curriculum in operation as "observable-and-reportable" (Garfinkel 1967), in other words that they learn to look and describe what goes on, not taking it for granted. On a more optimistic note, evaluation done from the start may also permit you to show at an early stage that you are achieving some of the specified results.

Who is to carry out the evaluation? The introduction of the KBSM has come from the top, so there is concern to make sure that the curriculum does not run into the possible problems already outlined. By implication then there is a need to develop a lower level focus for the implementation of the curriculum. The State Education officers as well as teachers have to feel that the curriculum is theirs and that they have a role in its developments. Eventually then it must be possible for people at all these levels to contribute to the project.

When is the evaluation to be done? This depends in part on having people interested in and trained to carry out evaluation. The commitment of Schools Division is to developing formative evaluation: the aim is to contribute to the improvement and modification of the curriculum implementation, the Division's responsibility. The evaluation findings will be there to help and advise (King, Morris and Fitz-Gibbon 1987). This means that work needs to start soon before people have become fixed in their attitude towards KBSM, and before any problems become entrenched. Formative evaluation needs to be a steady, continuous process, and will be complemented at intervals by findings from summative assessment. This does not mean that everyone will be doing evaluation all the time; introducing evaluation represents a cost, in terms of time if nothing else. The cost needs to be offset against the return, the payoff of the findings. But how are the findings to be obtained?

In looking at the how of this evaluation several issues come up. It is a collaborative enterprise, which means that while people must take responsibility for its findings, evaluation cannot be judgmental in the way that inspection is. Accountability implies that you have agreed the goals and will also agree when they have not been met. If evaluation is judgmental then confidence between people working at different levels will be lost, at a time when openness is essential to ensure that the scheme operates.

It was implied that the evaluation had to start as soon as possible. At the same time, when people are learning how to do evaluation, they need to take the work steadily, seeing how they can adapt to include it in their timetable. You cannot do everything at once, either, so have to focus on an issue that is important to you. There is little point to asking a question to which you do not want the answer, or where either way you will not be able to act on the information you glean. The opposition once held to exist between qualitative and quantitative approaches has been replaced by a view that they are complementary and even overlap (van Lier 1988). Both seem to be required in the kind of educational research evaluation is, and there are ways of establishing validity and reliability for both. The value of triangulation, whether of method or perspective, has to be understood, as do sampling techniques. However, what is certain is that people find it easier to start doing evaluation using techniques that do not require what most perceive as complex, even forbidding, statistical procedures!

Undertaking evaluation in the terms described here represents a particular approach to curriculum development. When, for example, officials form the State Language Office or the Ministry come to ask questions of teachers in schoo' a dialogue is being opened up. Asking a group of teachers how, for example, they could implement the curriculum more effectively suggests that their views will be heard. They will recognise that some of their requests cannot be met in a world which is not ideal, but they can still expect that realistic ideas will be attended to. A dialogue is being established then, where before there may not have been one, because evaluation depends on a two-way flow of information. Those taking part in the dialogue from below must feel that their views translate into action, and help to produce change. An important part of the management of innovation is the creation and maintenance of dialogue.

Putting the principles into practice

In this section I want to describe the initial stages of the project as seen by the project consultant. They are the first two phases covering the first nine months of the project in 1989. Three further phases are planned, taking the project to early 1991, when the first participants should begin the phase of inducting and training teachers to participate in the evaluation. The first five phases then are for the group of State and Ministry officials to gain experience and skills.

Practice: Phase One

The first phase began with a seminar in Melaka. After an introduction to the principles and some techniques of evaluation, the participants identified priority areas for evaluation of the implementation. Then n small groups they prepared sample questionnaires and observation schedules, which were reviewed by the plenum. They took part in a simulation on presenting evaluation to teachers, surprising themselves with the vehemence of some of the teachers' views. Finally, grouped according to their State or Ministry department, participants planned their intended evaluation and recorded this on two copies of an "Evaluation Proposal Form", one they kept and the other was given to the representatives of Schools Division to create a central record.

Overall in their end of course evaluation participants indicated that they had gained a good introduction to evaluation and how to plan and prepare for incorporating it into their work. A few showed they were aware that we had not gone through the whole process in detail. Most left feeling that evaluation would help develop the ELT programme. No overall fixed scheme for evaluation was proposed; it was felt that the participants needed some time in the field to explore the concept and learn how to evaluate. It seemed most useful that they should start with areas that they considered crucial, interesting or puzzling. Throughout the seminar it was emphasised that evaluation is investigative, collaborative, done at all levels of the system, and that it is dependent on a two-way flow of information.

Following the seminar it was possible for the consultant and a member of Schools Division to visit one State on the East Coast and one on the West of Peninsular Malaysia. During this fieldwork carried out with the State officials, including those who had attended the seminar, they visited eight schools to observe KBSM classes and to meet teachers. In discussion it was agreed that the first group should focus on one of the following areas identified during the visit:

(a) investigate the pupils' generally low motivation, particularly in rural schools; focussing on aspects of their attitudes towards English, knowledge of the English-speaking world and the place of English as an international language, and of proficiency in English as a requirement for employment;

(b) try out and evaluate techniques to use with slow learners;

(c) assess teachers' understanding and use of methodology required by or appropriate to changes introduced by KBSM (eg pair and group work; integration of skills; teaching moral values);

(d) evaluate locally-produced materials. The areas for the second group to consider were essentially the same as (a) and (c) above with the addition of:

(e) assess the adequacy of the briefing teachers have been given on KBSM;

The fieldwork provided an opportunity to come to more informed decisions about appropriate action; they revealed certain problems that had gone unnoticed; and it was possible to discuss practicalities in context. The officers all realised that the evaluation could not be rushed and that results would have to be worked for over a period of time. These visits are exactly the kind of support which should follow any such introductory course.

Practice: Phase Two

The second stage also began with a seminar, this time in Penang. Participants included just over half of the Melaka group and almost as many newcomers. The aims of the seminar were to:

(i) report on evaluation carried out in the States;

(ii) review principles and techniques of evaluation;

(iii) focus on the stages of analysis, interpretation and reporting of evaluation findings

There were practical sessions on analysis and reporting, as well as on techniques such as interviewing, observation, diary-studies, and case-study. Participants went through a simulated interview which they reported and then commented on. In State teams they started planning the evaluation work they could carry out over the next six months, and were urged to make this a small-scale, investigatory case-study; suggested areas for this were (a) teacher attitude to KBSM, (b) teaching in KBSM classes, and (c) pupil behaviour in KBSR and KBSM classes.

Firstly, the reports from the different state groups revealed a range of effort and experience. Findings were most interesting and useful where preparation and planning had been thought through carefully. One large scale survey had been successfully completed but had required a big commitment and use of free time by the team conducting it. In general, problems had arisen where aims had not been clearly enough defined, and where the instrument used had only been modified from the exercises carried out in Melaka, or had not been piloted. A few participants had been discouraged by their experience, and many were uncertain about how to report their findings: but this stage of evaluation was a major topic of this seminar. Overall the group's experience was positive and useful; given that they had not had any follow-up or assistance (with the exception, as it happened, of the large scale survey), their achievements were all the more satisfying as they had proved their independence.

However, participants left Penang knowing that they would get back-up they had not before; that there would be a newsletter to keep them in touch; that they had a limited objective for their next evaluation work and a focussed plan with a deadline for its achievement. The initial tendency of many in the group to want to be spoonfed with mechanical procedures had largely gone, though some were still not at ease with discovery learning! The seminar was able to build on and exploit their recent experience; their expectations became realistic as did their understanding of the objectives and process of evaluation. Two of the central States were visited for the fieldwork immediately following the seminar. The sample of ten schools represented a good range: urban, semi-urban and rural; single-sex, boys' and girls' schools as well as co-educational schools. The class visits were useful in developing a picture of teachers at work and the variety of conditions they have to meet. The urban-rural contrast is apparent as an underlying factor behind different levels of achievement, in favour of urban schools. However, the picture is modified by the success that can be achieved in a small, well-run school with enthusiastic teachers. Socio-enocomic differences, particularly in the semi-urban schools can influence performance in much the same way as the rural background does. These visits confirmed the earlier idea that an investigation of this variation in pupil performance and in possibly related pupil attitude might provide detailed understanding of something that potentially influences the way English in KBSM is received and can be taught.

In five group interviews teachers were asked to identify successes and problems they had in working with KBSM, and to suggest ways that they could improve their work. On the positive side teachers mentioned pupils who, after studying under the KBSR (Primary) curriculum, were more confident than their predecessors, and more fluent speakers of English; they also reported that with KBSM their classes were more interesting. The problems they reported were more numerous: KBSM brings an increased workload in preparation and administration: attention to Fluency seems to bring with it a decrease in Accuracy; there are difficulties with integration of skills and of moral values; there seems to be excessive emphasis on phonology in the curriculum and in textbooks; teachers would like to be able to choose textbooks as some of those on offer are boring, underestimate pupils and lack a range of exercises.

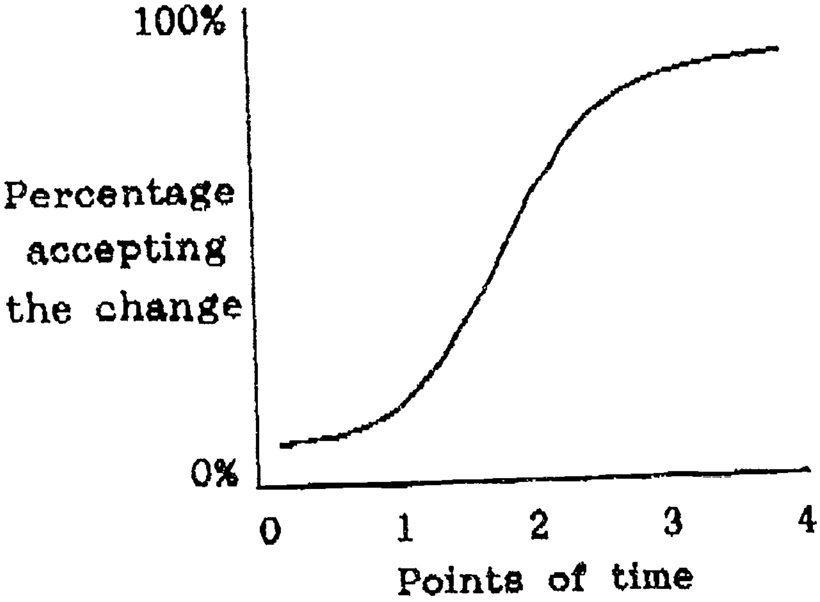
Training courses for KBSM were reported to have been too theoretical in the first instance, but later courses had provided plenty of practical guidance. Many teachers are not confident that they are doing what is required. They are not sure how to select topics and activities or how to adapt them to their pupils. KBSM in the view of several teachers was suitable for better learners from privilege backgrounds, but not for slow learners from rural or low-income families.

Set against these difficulties, and accompanying requests for help, some teachers provided models of appropriate, independent action: a group of teachers in one school who worked together to produce resource material and bank it; in another school the group coordinated their work and consulted each other; another group were developing appropriate new tests in the absence of a central model; and some teachers did not allow the curriculum to dictate their work, focusing instead on their learners, making an appropriate interpretation of the curriculum for their audience. One teacher in a rural school carried out evaluation of her performance with her pupils. In effect the innovation is already under way at the bottom, on a limited scale which up till now has lacked support; now it can be given more direction and other teachers can hear about it.

CONCLUSION

At one level evaluation aims to channel teachers' energy from inactive preoccupation with their anxieties and difficulties to seeking solutions to their problems, evaluation calls for greater involvement in their work and offers them the chance to improve their professional skills. On another level it aims to guide officials in making decisions, developing the efficiency and effectiveness of the curriculum, materials and teaching, and choosing appropriate support for teachers through in-service training.

None of this can happen until there is general acceptance of the potential value of doing evaluation, then training to carry it out, before slowly gaining experience and expertise. Innovations of this kind usually build up their acceptance in an S-curve (Fig 1; cf Markee 1990, White 1988):



The current project seems to have a large group of "early adopters" (cf White 1988) if we judge by the end of seminar returns; those leading the innovation will need to find other ways to measure adoption than through expressions of faith at this point and in a questionnaire. One way is through the evidence of adoption from planning, the creation of appropriate instruments, and the quality of reporting. These criteria would bring the number of adopters at the end of phase one to a more probable but still sizeable proportion.

It is much too soon to estimate the effect of this project. In planning this innovation particular attention has been paid to establishing a network of innovators, aiming to develop confidence in their new role through experience; to the need for widespread communication; to the collaborative, responsible nature of the enterprise; to identifying existing practice where staff meetings to discuss problems and new methods could with a little encouragement and guidance become more effective fora for staff development and the gathering of useful monitored information. The project has central support, and the State Offices receive visits from Schools Division; there is a newsletter and participants meet locally and nationally; and reference material has been provided by the British Council.

In the last few months the press in Malaysia has paid considerable attention to official and public concern about the teaching of English at secondary level. This has been prompted by a lack of suitably qualified teachers, concern about standards of achievement, and by a general need to raise public awareness of the importance of the language for international use in trade and diplomacy. This project has the potential to make a distinctive contribution to meeting these needs.

ACKNOWLEDGEMENTS

I would like to thank the Schools Division of the Malaysian Ministry of Education and the British Council, Malaysia for permission to report on this work. I also want to thank the participants in the seminars, as well as the teachers and pupils who let me into their classes and then took time to talk about their work, for their help and stimulation in carrying out the assignments on which this paper is based.

BIBLIOGRAPHY

Alderson J C (ed) 1985 Evaluation: Lancaster Practical Papers in English Language Education vol 6 Pergamon.

Brindley G 1989 Assessing Achievement in the Learner-Centred Curriculum NCELTR

Corder S P 1968 "Advanced study and the experienced teacher" in Perren G E (ed) Teachers of English as a Second Language Cambridge University Press.

Garfinkel H 1967 Studies in Ethnomethodology Prentice-Hall.

Goh K S et al 1989 Compendium: a handbook for ELT teachers Kementerian Pendidikan Malaysia.

van Lier L 1988 The Classroom and the Language Learner Longman White R V 1988 The ELT Curriculum Blackwell.

Holt M 1987 Judgment, Planning and Educational Change Harper and Row.

King J A, L L Morris and C T Fitz-Gibbon 1987 How to Assess Program Implementation Sage Publications.

Markee N 1990 "The diffusion of communicative innovations and classroom culture: an ethnographic study" paper presented at TESOL Annual Convention, San Francisco C A.

Murphy D F 1981 "Developing secondary teachers' evaluation of their work" in Practical Papers in English Language Education vol 4 University of Lancaster 1985 "Evaluation in language teaching: assessment, accountability and awareness" in Alderson J C (ed).

Nunan D 1988 The Learner-Centred Curriculum University Press van Lier L 1988 The Classroom and the Language Learner Longman White R V 1988 The ELT Curriculum Blackwell.

NATIONAL LEVEL FORMATIVE EVALUATION: SOME FIRST STEPS

Ali Abdul Ghani and Brian Hunt

1 History and Background

From the beginning of the academic year in January 1987 a new integrated language curriculum: Kurikulum Bersepadu Sekolah Menengah (KBSM), was introduced into Malaysian secondary schools. This was for all the languages taught in Malaysia at secondary school level: Bahasa Malaysia, English, Chinese, Tamil. The KBSM was to be introduced progressively year by year beginning with Form I in 1988. 1990 sees its introduction into Form III. The introduction of the KBSM curriculum for other subjects is following one year behind.

The KBSM curriculum for English aims to integrate the language skills (listening, speaking, reading and writing); the language areas of lexis, phonology and grammar as well as knowledge of other subjects on the timetable and good moral values as indicated by the National Education Philosophy of Malaysia.

The Ministry of Education felt that the introduction of the KBSM curriculum should be monitored in order to gauge its effects on teaching and learning in the classroom. It was felt that teachers and students, as the ultimate users of the new curriculum, would have most to say about its effectiveness.

There are three reasons for this decision.

Firstly, it is hoped to obtain a clear picture of the effects of the new curriculum on classroom teaching and learning.

Secondly, it is hoped that teacher-generated ideas for curriculum development will be more sensitive to learners' needs.

Thirdly, modifications to the syllabus are likely to be longer lasting if they are recommended and carried out from 'bottom up' rather than prescribed from 'top down'.

Having decided, in principle, to monitor the introduction of the new curriculum a choice had to be made between a summative evaluation and a formation evaluation.

Several reasons decided in favour of a formative, rather than a summative, evaluation.

Firstly, as the KBSM curriculum is to be introduced progressively year-by-year beginning with the first year and ending with fifth year of secondary schooling, conducting a summative evaluation exercise at the end of five years would be too late to be of optimum benefit.

Secondly, in addition to being too late, it was felt that a summative evaluation would be less subtle and sensitive for the task of monitoring the KBSM programme; and that the ultimate aim of the formative evaluation exercise should be the creation of a delicate set of evaluation instruments.

Dermot Murphy explicitly deals with this aspect:

> 'Summative evaluation, most often realized as assessment of learner performance, can produce results open to interpretation as dealing with something fixed: what the results should be. Poor results are due to the learners not achieving the objectives of the course. ***Its focus is limited, and the assessment may not give any clue to what needs adjusting to match learner achievement to curriculum expectation.*** The evaluation does

not produce enough information and reinforces professional secrecy and entrenchment'. (Our emphasis)

Dermot Murphy 'Evaluation in Language Teaching: Assessment, Accountability and Awareness'. Lancaster Practical Papers in English Language Education. Volume 6. Pergamon Press (1985)

Thirdly, the Ministry is eager to encourage teachers to contribute their ideas to curriculum evaluation and development, and by initiating a 'bottom-up' formative evaluation project it is hoped to engage in valuable and informative dialogues between teachers and the Ministry of Education.

Fourthly, it is felt that by encouraging classroom-based investigations of the curriculum in use by teachers and learners, teachers can develop their own self awareness of their classroom techniques and methodology with a view to professional development in both the long and short terms.

Another important consideration relates to the separate responsibilities of the two divisions of the Ministry whose work relates to schools and curricula. Schools Division has the responsibility, through State Education Offices, for the administration and management of schools, teachers and students. The Curriculum Development Centre has the responsibility for curriculum policy, planning, the introduction of new curricula and related training programmes for Resource Persons. It is important that any programme of monitoring the KBSM does not duplicate the work of other ministerial divisions. Diverse divisions of the ministry are, however, kept informed of English language programmes via periodic meetings of inter-divisional committees. The progress and interim findings of the formative evaluation programme are thus reported regularly.

We hope that the results of a formative evaluation will,

i provide precise & accurate information to relevant divisions of the Ministry about the existing situation of teaching and learning in Malaysian secondary schools.

ii allow refinements to be made to the curriculum in the light of experience.

iii help in the planning of teacher education courses related to the KBSM curriculum.

2 Implementation of the Programme

The former English Language Adviser to Schools Division, Mr. Lionel Thompson, was asked to investigate the possibility of an evaluation project. He prepared a working paper in which he outlined areas for investigation. These included suggested investigation of systemic, environmental and pedagogic factors. An initial approach was made on paper to Mr. Dermot Murphy,lecturer in ELT at St. Mary' College, Twickenham, whose response was enthusiastic.

To initiate the formative evaluation project Dermot Murphy was invited to Malaysia to conduct a one-week introductory training seminar, and to follow up with practical field training lasting a further week.

Forty-four participants consisting of State Education Language Officers, Key Personnel and Resource Personnel from each of the fourteen states; officers from various interested divisions of the Ministry of Education (including Heads of Bahasa Malaysia, Chinese and Tamil), Curriculum Development Centre, Federal Inspectorate, Teacher Education Division, and Examinations Syndicate attended the seminar held in Melaka.

The seminar introduced participants to the principles and purpose of evaluation; information gathering techniques and sampling. A range of evaluation questionnaires and pro-forma were analysed. Participants discussed areas for evaluation, and identified a range of priorities. The areas for investigation suggested by Mr. Thompson were offered as a starting point.

The participants prepared evaluation instruments, presented their ideas to fellow-participants and refined them in group discussion. As an aid to classroom observation techniques a video from 'Teaching and Learning in Focus' was presented, and discussed. A simulation was prepared and carried out so that participants could, in turn, present the notion of formative evaluation to teachers. Finally, participants organized and planned future work areas based on ideas gleaned from the seminar.

During the practical field visits to Kelantan (the state in the extreme north-east of Peninsula Malaysia) and Melaka (the state in the south-west of Peninsula Malaysia), State Education Officers, in consultation with Dermot Murphy and school teachers, agreed on their priorities for investigation.

In Kelantan these were:

a investigate students' generally low motivation, particularly in rural schools, focussing on aspects of their attitudes towards English; knowledge of the English-speaking world; the place of English as an international language; and of proficiency in English as a requirement for employment;

b try out and evaluate teaching techniques to use with slow learners;

c assess teachers' understanding and use of methodology required by, or appropriate to, changes introduced by KBSM (e.g. pair and group work; integration of skills; integration of moral values);

d evaluate locally produced materials;

In Melaka these were:

a investigate motivation (as in Kelantan); examine in particular the lack of success a particular semi-rural school has had with the English Language Reading Programme;

b assess the adequacy of the briefing teachers have been given on KBSM;

c assess teachers' understanding and use of methodology required by or appropriate to changes introduced by KBSM (e.g. pair and group work; integration of skills; integration moral values).

It was proposed at the Melaka seminar that participants would organize their own exploratory investigations and that a future seminar to allow reporting back would be arranged. Accordingly Mr. Dermot Murphy was reinvited to conduct a seminar in Penang in September 1989.

Forty-four participants attended the Penang seminar. Participants were State Language Officers, Supervisors, Key Personnel, Resource Personnel, Inspectors and one teacher. This number included twenty-four officers who had attended the Melaka seminar who were directly involved in formative evaluation work, plus colleagues who had been introduced to the formative evaluation programme following the Melaka seminar.

The aims of the seminar were to

i report on evaluation carried out thus far in the individual States

ii review principles and techniques for evaluation

iii focus on the stages of analysis, interpretation and reporting of evaluation findings

iv plan for the third stage of the project.

Participants reported back on their work to date and explained how their investigative instruments had been designed. Many participants felt that they had taken on too much work.

Practical work on analysis and reporting was carried out from participants' work. Dermot Murphy reviewed a range of data gathering techniques: interviewing, observation, diary studies and case-study, to allow participants opportunities to 'triangulate' from their existing data.

By the end of the seminar participants were able to focus their future work. Suggested areas were

i teacher attitude to KBSM

ii teaching techniques in KBSM classes

iii pupil behaviour in KBSR (primary) and KBSM (secondary) classes.

3 The Formative Evaluation Project in Operation

There are several strands of personnel in our formative evaluation project at the Ministry of Education in Malaysia. These are:

Personnel	Roles
Mr Dermot Murphy St. Mary's College Twickenham	act as consultant to the programme provide academic input and practical training (through seminars).
Ministry of Education Officers	manage overall project provide support and guidance keep states up-to-date and in touch with each other.
State Language Officers from all 14 states Key Personnel and Resource Persons from all states	design evaluation instruments collect data.
Teachers in various schools Students in various schools	provide data and feedback from personal experience.

From the inception of his association with our formative evaluation project Mr. Dermot Murphy has had three main tasks.

Firstly he was to provide information about formative evaluation to officers involved in the project. Secondly to give training in preparing evaluation instruments, data collection, data interpretation and reporting. He did all this and more during our first two seminars and will continue with this work during our forthcoming seminar in July, 1990.

Thirdly he was to make recommendations to the Ministry of Education concerning the development of the project in the longer term. His reports have been most useful and our project continues to benefit greatly from his advice. We are working closely with Mr. Murphy in respect of the stages in the project and in the direction that the project should take.

The task of the Schools Division of the Ministry of Education is to oversee the development of the project from its inception to becoming a part of our educational administrative system.

We also have the responsibility to give support to the teams working on formative evaluation at state level through:

- a quarterly newsletter (to keep states in touch and up-to-date)
- visit to states (to give assistance and advise where needed)
- organizing regular (usually twice a year) meetings with all state language officers (to review progress).

Following Mr. Murphy's recommendations we have divided our formative evaluation project into five phases lasting over some 2 1/2 years from January 1989 until June 1991.

In general, these phases cover the stages of development of the project from the initial introduction to the project in January 1989 until the submission of the first set of interpreted data along with a report of related recommendations to senior Ministry of Education officials in May 1991. The time framework is to a certain extent prescribed by the support and funding by the Overseas Development Agency, London of an ELT adviser to the Schools Division of the Ministry of Education.

(Details of our project stages are in appendix I).

Areas of Investigation

Within broad guidelines (e.g. the desirability of teacher-led, classroom-based research; the need to monitor the effectiveness of the KBSM curriculum) State Language Officers and their teams were purposely allowed a free hand in deciding their topics for investigation, (in collaboration with teachers), designing their investigative instruments and for drawing up their own administrative schedules. Although there is predictably a degree of overlap in the areas chosen for investigation, there are a range of differences as well as variety of investigative techniques.

The most common topic is some aspect of teachers' classroom management and methodology with particular reference to integration of skills and content, small group activities, pairwork, use of teaching materials and other teaching aids.

Ten out of the fourteen states are investigating one or more aspects of classroom management and methodolcgy. One in-depth study is concentrating on the classroom methodologies of teachers wio have little or no ELT training but who have been seconded to teach ELT because of their English language proficiency. This investigation is using questionnaires and interviews, lesson observation, self-evaluation forms and diary studies.

A smaller number of states (four in all) have chosen to direct their investigations towards learning. Specific areas of research are class participation, student interaction, learning styles, story telling and language games, and students' responses to different types of homework assignment. This research is being carried out through the use of questionnaires and interviews (with both teachers and students).

What will prove useful for our national level investigations is that different investigative tools have been chosen by the various states. This means that classroom management and methodology is being investigated using questionnaires and interviews (with both teachers and students); lesson observations, teacher self-assessment forms and teachers end-of-class checklists. As Dermot Murphy has repeatedly stressed the notion of triangulation during each of the seminars held so far, the fact that we will be able to select from a range of data collection techniques for our national level investigations will allow us to scrutinize our data gathering carefully and be more confident in the validity of our results.

(See Appendices II and III for a fuller outline of areas of investigation, and Appendix IV for an outline of projects state by state).

What Have We Learned So Far?

While preparing our paper for this conference we held a two-day meeting with our State Language Officers during which we asked them to give us an up-date of their projects with particular reference to any advice they could now offer, from personal experience, to those contemplating formative evaluation research projects.

This is the advice they offer, more or less unanimously.

Planning	- plan carefully - prepare appropriate instruments - get a third opinion on your instruments - conduct a small pilot study to ensure validity of instruments and whether users can understand it. - try not to do too much!
Topics	- investigate one aspect at a time - identify one area for analysis and investigate this thoroughly (NB triangulation) - start small to begin with.
Questionnaires	- precise, clear, with specific objective - students' questionnaires should be bilingual (Bahasa Malaysia & English)
Personnel	- do not involve too many people initially - get a good team! - get people who are willing to carry out their tasks.
Reporting	- spend time and effort refining your system of reporting (it will save you both time and effort later!).
Public Relations	- try and involve other subjects and other interested parties publicize your efforts.

Participants' responses

At the Ministry we were concerned that the formative evaluation project would stimulate our personnel to initiate research projects. We were particularly anxious that State Language Officers and their teams, and teachers who were involved in the project would not be so pressured by other priorities that they would be unwilling to conduct time-consuming investigations of the type we envisaged.

After the introductory seminar in Melaka in January 1989 we learnt from informal feedback that a number of participants had reservations about the utility of the project or indeed about the Ministry's sincerity in engaging in a dialogue with classroom practitioners. A proportion felt that the project would be short-lived.

On the practical side participants felt unsure and perhaps over-cautions about what to do next. Some felt that they had not been given enough support during the seminar (although this was one aim of the 'discovery' learning approach).

Nevertheless, state level teams suspended their disbelief and organized working committees, produced research tools and began their investigations. Although results so far have been incomplete and inconclusive, some of them are summarized below.

Group Work

i Teachers are concerned that the noise level in their classes during groupwork may 'upset' their Principals.

ii The good students contribute most of the work during groupwork; the weaker students remain silent.

iii There is a tendency during groupwork for students to use Bahasa Malaysia or their mother tongue.

Integration of Skills

i Teachers feel that more attention should first be given to oral skills.

ii Students are reluctant to speak.

iii There is insufficient time in the class to integrate all language skills.

Moral Values

i It is difficult to include moral values in all lessons.

ii Sometimes lessons may become boring because of attempts to incorporate moral values.

What we have learned at the Ministry is:

Rome was not built in a day

In the interests of having a firm base of experienced and trained personnel we are prepared to allow time for projects to be trialled, refined and tried again. We think that this will allow our officers to gain expertise from personal experience.

More haste less speed.

We regard the formative evaluation work as a long term, ongoing part of the administration of our education system. We believe that rushing things at the beginning of the project, and we feel that we are still at the beginning of our work, will only create problems for us at a later date.

We have tried to give an overview of our formative evaluation project which, aiming towards a national level investigation, naturally involves many people from different areas of our education network. Perhaps the most important thing we have learned is that evaluation of educational processes is, like education itself, continuous and long term.

Appendix I: Project Stages

The stages of our project are:

STAGE	CONTENT
phase one Melaka Seminar January 1989	Introduce personnel to the aims and benefits of formative evaluation. Demonstrate types of data collection. Instigate first attempts.
phase two Penang Seminar August 1989	Review first attempts. Introduce a quarterly newsletter. Schools Division to visit each state to monitor work and 'fine tune' instruments.
phase three Genting Seminar July 1990	Review Selection of state level instruments for national level work. of existing instruments. Framework of channels of reporting from school level via state level to ministry level. Prepare administrative network (e.g. job descriptions). Introduce personnel to quantitative procedures, and train personnel in more sophisticated ways of analysis and reporting. Instigate National level investigations.
phase four December 1990	Review of progress and available instruments. Seminar and workshops to introduce formative evaluation work into schools nationally. Schools Division to visit each state to help with data collection
phase five	Interpretation of first set of March 1991 findings by Schools Division. Preparation of report on initial national formative evaluation. Survey and related recommendations for submission to Ministry officials.

Appendix II: Areas of investigation

Focusing on Methodology

teachers' methodology	questionnaires (teachers & students) class observation interviews pro-forms
classroom management	observations questionnaires
managing group activities	questionnaires interviews (students & teachers)
managing pair work	questionnaires interviews
integration of skills & content	questionnaires interviews
use of teaching aids and materials	questionnaires checklists interviews (teachers)
investigation of teaching and learn- learning techniques in the KBSM syllabus	questionnaires
a study of non-optionists (ELT teachers whose main discipline is not English)	observation self-evaluation forms. questionnaires interviews diary studies

Appendix III : Areas of investigation

Focusing on Learning

class participation/student interaction	questionnaires and interviews (students & teachers)
story telling & language games	questionnaires (students)
students' work activities	student questionnaires
the level of class participation	questionnaires and interviews (teachers & students)

Appendix IV: State Formative Evaluation Projects

Formative Evaluation projects are being conducted in each state in Malaysia.

State	Personnel	Number of schools	Area of investigation	Data collection technique(s)
Perlis	2 JPN officers 8 teachers questionnaires	16 secondary schools	teaching techniques	questionnaires (students (teachers) lesson observations
Keda'n	2 JPN officers 3 teachers	30 secondary schools	students homework	questionnaires (forms I & II assignments students)
Pulau Pinang	2 JPN officers 1 Key Personnel 1 teacher	20 secondary schools	students' class (10 urban) (10 rural)	questionnaires interviews participation students class interaction
Perak	3 JPN officers 12 Resource Persons 9 Assistant District Education officers	20 secondary schools of teaching	teachers' checklist aids and teaching materials	use questionnaires
Kelantan	3 JPN officers 20 teachers	20 secondary schools	non-optionists using KBSM	lesson observations checklists self-evaluation forms questionnaires in reviews diary studies
Terengganu	1 JPN officer teachers	11 secondary schools	teaching methodologies	questicnnaires interviews 10 State and District
Pahang	2 JPN officers 12 Key Personnel	15 secondary schools	teaching methodologies in KBSM learning opportunities in KBSM	questionnaires

State	Personnel	Number of schools	Area of investigation	Data collection technique(s)
Wilayah Persekutuan	3 JPN officers 2 Key Persons Supervisor 3 teachers	15 secondary schools	story telling and language	questionnaires 1 English games
Selangor	2 JPN officers 3 Resource Persons 1 Language Supervisor teachers students	15 secondary schools	pairwork activities	questionnaires interviews
Johor	1 JPN officer 17 Key Persons	142 secondary schools (both rural and urban)	groupwork activities	questionnaires lesson observations
Negeri Sembilan	4 JPN officers 3 Resource	10 secondary schools	classroom management	observations questionnaires
Melaka	4 JPN officers 3 Resource Persons	12 secondary schools	integration of skills & content	questionnaires (students & teachers) interviews small group (students and activities teachers) moral values
Sarawak	2 JPN officers 1 Assistant Principal 4 teachers	12 secondary schools	managing group activities	questionnaires interviews
Sabah	2 JPN officers School Inspectors Zone Heads Principals Resource Persons Teachers	10 secondary schools	teaching methodology	questionnaires interviews proforma

SECOND LANGUAGE PROFICIENCY ASSESSMENT AND PROGRAM EVALUATION

David Nunan

INTRODUCTION

I have been asked, today, to examine the role of second language proficiency assessment in program evaluation. In the paper, I shall argue that while assessment is an important component of program evaluation, it is only one component. I shall further argue against the construct of general 'curriculum-free' proficiency, as this is currently operationalized in the literature, as a central component in program evaluation. 'Curriculum-free' proficiency is proficiency which is not tied to or referenced against curriculum goals. My reservations about the use of 'proficiency', thus conceived, as a central element in program evaluation are four in number, and will be expanded upon in the course of the presentation.

1 The construct of proficiency has not been operationalized in a way which enables it to be usefully used for the purposes of program evaluation.

2 Criterion-referenced measures of achievement are of more practical utility than statements of proficiency which are not related to program goals.

3 Regardless of the terms in which learner outcomes are to be defined, comprehensive program evaluation requires the collection, interpretation and evaluation of data relating to a range of processes and elements operating within a particular educational context, not just learner outcomes.

4 In order to interpret outcome data, one needs process data.

The paper contains a number of practical suggestions which have implications for carrying out program evaluation within a Southeast Asian context, and includes some sample instruments for carrying out such evaluations.

THE CONCEPTS OF LANGUAGE PROFICIENCY AND EVALUATION

This paper is centrally concerned with proficiency assessment and evaluation, and I should therefore attempt to clarify my understanding of these terms from the outset. In some educational systems, the terms 'assessment' and 'evaluation' are used interchangeably - witness the following quote from Gronlund:

> Evaluation may be defined as a systematic process of determining the extent to which instructional objectives are achieved by pupils. There are two important aspects of this definition. First, note that evaluation implies a systematic process, which limits casual, uncontrolled observation of pupils. Second, evaluation assumes that instructional objectives have been previously identified. Without previously determined objectives, it is difficult to judge clearly the nature and extent of pupil learning.
>
> (Gronlund 1981:5)

Gronlund, in circumscribing evaluation in terms of learning outcomes, presents an extremely narrow input-output view of evaluation and, by extension, education. In fact, he is using the term 'evaluation' roughly in the sense in which I would use 'assessment'. I would like to suggest that, while they are obviously related, they mean rather different things. To me, assessment refers to the set of processes through which we make judgements about what a learner is able to do in the target language. We may or may not assume that such abilities have been brought about by a program of study.

'Evaluation' is a wider term than 'assessment'. While it entails the collection of information on what learners can do in the target language it also involves additional processes designed to assist us in interpreting and acting on the results of our assessment. The data resulting from evaluation assist us in deciding whether a course needs to be modified or altered in any way so that objectives can be achieved more effectively. If certain learners are not achieving the goals and objectives set for a course, it is necessary to determine why this is so. We would also wish, as a result of evaluating a course, to have some idea about what measures might be taken to remedy any shortcomings. Evaluation, then, is not simply a process of obtaining information, it is also a decision-making process.

In this area, there seems to be a certain tension between 'measurement' and 'evaluation'. Those who are seduced by the illusion of certainty offered by tools and techniques for measuring things sometimes seem to forget that there is an essential difference between the value neutral processes of measurement and the value laden nature of evaluation (Wolf 1984).

Thus far, I have argued that assessment is a process of collecting information about what a learner can do in the target language, while program evaluation is a more general process of obtaining a variety of information relating to different curriculum elements and processes, for decision-making purposes. For most evaluations, I believe it is useful to collect data on what learners can and cannot do, although this view is by no means universally held by program evaluators, and for some types of evaluation it may be either unnecessary or impossible to obtain such data.

In recent years, a great deal has been written and said about the use of measures of proficiency as a means of assessing learners. I believe that there are some serious problems with the way the concept of proficiency has been defined and operationalised, and in this section I shall explore some of these problems. This will provide a basis for considering the feasibility or desirability of adopting a 'program-free'approach to proficiency assessment. Before we consider assessment instruments themselves, however, it is necessary to engage in some terminological ground clearing.

Within the literature, there is considerable confusion about the constructs and terminology associated with language development and use. Confusion, disagreement and uncertainty are reflected in much of the writing associated with language testing, a confusion which can be partly explained by a lack of agreement about the nature of language, language learning and use. This confusion is evident in the various ways in which terms such as 'competence', 'performance', 'proficiency' and so on are used. Although he did not create the terms, Chomsky (1965) gave prominence to the notions of 'competence' and 'performance'. For Chomsky, 'competence' refers to the mastery of principles governing language bahaviour. 'Performance, on the other hand, refers to the manifestation of these internalised rules in actual language use. The terms have come to be used to refer to what a person knows about a language (competence) in contrast to wha\ that person does (performance). More recently the term 'communicative competence' has gained currency, and there has been some debate as to the actual constituents of this construct. There is also considerable ongoing debate about what it means to 'know the rules of a given language'.

Diller (1978) attempts to resolve this paradox by suggesting that knowledge exists on a subconscious level:

> ... if children are not able to formulate the rules of grammar which they use, in what sense can we ,ay that they 'know' these rules? This is the question which has bothered linguists. The answer is that they know the rules in a functional way, in a way which relates the changes in abstract grammatical structure to changes in meaning. Knowledge does not always have to be consciously formulated. Children can use tools before they learn the names for these tools.
>
> (Diller 1978: 26-27)

If we accept that knowledge need not be consciously formulated, but may manifest itself in the ability to use the language, it would seem to render the competence-performance distinction rather uncertain. (See also the systemic-functionalist view that the distinction is unnecessary and misleading because language is what language does.)

Krashen (1981, 1982) further confuses the issue by suggesting that knowledge of linguistic rules is the outward manifestation of one psychological construct (learning), while use of these rules to communicate is the manifestation of another construct (acquisition).

Rea (1985) subsequently questioned the need for a 'competence' construct by suggesting that as we can only observe instances of performance, not competence, the competence-performance distinction is redundant. In testing terms, she suggests that we forget about 'competence' and think in terms of communicative performance and non-communicative performance.

This brings us to the point where linguistic knowledge is to be defined in terms of what an individual is able to do with that knowledge. This is reflected in the competency-based ESL movement which has gained a certain amount of promirence, particularly in the United States. As though there were not enough confusion over terminology, this movement is using 'competence' to refer to things learners can do with language; that is, it is used in roughly the same sense as 'performance' in the earlier competence-performance distinction. In ESL, 'a competency is a task-oriented goal written in terms of behavioural objectives' (CAL 1983:9) which has clear implications for assessment. Assessment is built in. Once the competency has been identified, it also serves as a means of evaluating student performance. Since it is performance based, assesment rests on whether the student can perform the competency or not. The only problem is to establish the level at which the student can perform the competency. (op cit:11-13)

Within the literature, some writers use the term 'proficiency' as an alternative to 'competency' (see, for example Higgs 1984). Richards, however, makes a clear distinction between 'competence' and 'proficiency', although he characterises the concept of proficiency in the same way as Competency Based Education characterises competency:

> 1 When we speak of proficiency, we are not referring to knowledge of a language, that is, to abstract, mental and unobservable abilities. We are referring to performance, or, that is, to observable or measurable behaviour. Whereas competence refers to what we know about the rules of use and rules of speaking of a language, proficiency refers to how well we can use such rules in communication.
>
> 2 Proficiency is always described in terms of real-world tasks, being defined with reference to specific situations settings purposes activities and so on.
>
> (Richards 1985: 5)

Richards goes on to argue that:

> A proficiency-oriented language curriculum is not one which sets out to teach learners linguistic or communicative competence, since these are merely abstractions or idealisations: rather, it is organised around the particular kinds of communicative tasks the learners need to master and the skills and bahaviours needed to accomplish them. The goal of a proficiency-based curriculum is not to provide opportunities for the learners to 'acquire' the target language: it is to enable learners to develop the skills needed to use language for specific purposes.
>
> (Richards 1985: 5)

In this section, I have attempted to highlight some of the confusion surrounding key concepts relating to the nature of language proficiency. This confusion is due partly to the inconsistent application of terms to concepts and partly to confusion over the nature of the concepts themselves. If we follow the portrayal of Richards, proficiency, simply put, refers to the ability to perform real-world tasks with a prespecified degree of skill. In programmatic terms this definition is probably reasonable enough. However, when it comes to the assessment of second language proficiency, the psychological reality of the construct become problematic, as we shall now see.

In order to assess any area of human behaviour, it is necessary to have some idea of what it is we are trying to assess. What is it that testers of language proficiency are trying to assess? We can get some idea by looking at the instruments which have been developed. One increasingly popular instrument is the proficiency rating scale. What follows is the generic description of speaking profieicny at an intermediate-high level. It is taken from the American Council on the Teaching of Foreign Languages Provisional Proficiency Guidelines.

Able to satisfy most survival needs and limited social demands.
Shows some sponteneity in language production but fluency is very uneven.
Can initiate and sustain a general conversation but has little understanding of the social conventions of conversation.
Developing flexibility in a range of sircumstances beyond immediate survival needs.
Limited vocabulary range necessitates much hesitation and circumlocution.
The commoner tense form occur but are frequent in formation and selection.
Can use most question forms.
While some word order is established, errors still occur in more complex patterns.
Cannot sustain coherent structures in longer utterances or unfamiliar situations.
Ability to describe and give precise information is limited.
Aware of basic cohesive features such as pronouns and verb inflections, but many are unreliable, especially if less immediate in reference.
Extended discourse is largely a series of short, discrete utterances.
Articulation is comprehensible to native speakers used to dealing with foreigners, and can combine most phonemes with reasonable comprehensibility, but still has difficulty in producing certain sounds in certain positions or in certain combinations, and speech will usually be laboured.
Still has to repeat utterances frequently to be understood by the general public.
Able to produce some narration in either past or future.
(Cited in Savignon and Berns 1984: 228-229)

The use of such scales is fraught with hidden dangers, which, for reasons of space, can only be briefly sketched out here. The scales themselves tend to take on ontological status - that is, there is a tendency to assume that such a person as an 'Intermediate-High' learner actually exists and that there is such a thing as 'Intermediate-High' ability - rather than being something constructed to account for observable or hypothetical features of learners' speech. (See also, Lantolf and Frawley, 1988 who point out the essential circularity of the descriptions). The scales themselves have not always been empirically validated to determine if learners really do act in the ways described by the scales. Research from second language language acquisition is often overlooked or ignored. (Some scales actually violate findings from SLA research.) One rating scale (the Australian Second Language Proficiency Rating Scale) makes claims about the equivalence of real world tasks and their appropriacy at different levels of proficiency. It is suggested, for example, that the tasks of 'returning an unsatisfactory purchase' and 'explaining some personal symptoms to a doctor' are of the same order of difficulty. However, no empirical evidence is provided that these tasks draw on the same linguistic and communicative resources, nor that the ability to perform such tasks can be determined by indirect measures of proficiency such as an oral interview. Finally, in terms of construct validity, the scales confound phonological, morphosyntactic, lexical, semantic and pragmatic features.

Program-free proficiency assessment and learner achievement

Within the literature, there are claims that program evaluation should be based on tests of general language proficiency through means such as the proficiency rating scales critiqued in the preceding section, not on achievement measures which are related to or associated with the program being evaluated. This line of argument is based on the view that unless transfer of learning can be demonstrated to have taken place, then learning, in any meaningful sense can not be said to have taken place. ('Transfer' is generally defined as the extent to which knowledge and skills developed in one field can be taught in a way which enables them to be utilized in another field.) There are a number of problems associated with the above argument, as we shall shortly see. In fact, even if learning transfer can be demonstrated to have occurred, it is quite another matter to demonstrate that learning is the result of a specific program intervention.

The whole issue of transfer of learning has, of course, been long debated in the educational and cognitive psychology literature. One debate concerns the relative claims of the cognitive skills transfer hypothesis versus the subject-domain hypothesis. The cognitive skills transfer hypothesis suggests that the development of knowledge and skills in certain subject domains can develop general learning and thinking skills which will transfer to other subject domains. For example, in a Western context, the teaching of languages, particularly Latin and Greek, was, for many years, defended on the grounds that it facilitated the development of reasoning skills which could be subsequently employed on more relevant subject areas. However, there has never been any evidence to support this claim. In fact, what evidence there is seems to run counter to the claim (see, for example, Thorndike and Woodward 1981, and Resnick 1987 cited in De Corte 1987). In contrast to the paucity of data on the transferability of general learning skills, there is a great deal of evidence to suggest that "the availability and flexible use of a well-ordered body of domain-specific knowledge play a major role in successful learning and problem-solving activities." (Glaser 1987).

Voss (1987) provides a reconceptualisation of the concepts of learning and transfer based upon a general information processing model of problem solving which suggests that learning and acquisition are subordinate to transfer. His paper begins with an analysis of the concepts 'acquisition', 'learning' and 'transfer', as defined by Association Theory which derived its definitions from everyday knowledge rather than systematic analysis. 'Acquisition' was investigated in "multiple trila experiments which intrinsically presumed contiguity and frequency as the mechanisms producing acquisition". 'Learning' was defined as an improvement in performance as a result of practice, while 'transfer' was defined as "the influence of the learning of one task upon the performance of a second task" (Voss 1987: 608). With the demise of associationism came a decrease in the use of multiple trial acquisition experiments and the use of the concepts 'learning', 'retention' and 'transfer'.

Voss outlines Jenkins' tetrahedal model which suggests that learning and memory are dependent on the interaction between four classes of variables. These are 'orienting task' (e.g. instructions, activities); materials (e.g. sensory mode, physical structure); criterial tasks (e.g. recall, recognition, problem-solving); subject characteristics (e.g. activities, interests, knowledge). As the manipulation of two or more of these variables results in a significant interaction, it is almost impossible to conduct laboratory experiments which will yield generalisable results. The thrust of Jenkins' work is to suggest that:

> ... there is no one way to learn since learning wil depend on the instructional task, the materials, the criterion of learning and the characteristics of the individual who is learning. The answer to the question of how best to teach a particular subject matter to a particular group of subjects becomes "it depends".
>
> (Voss 1987: 609)

Given these criticisms, Voss sets out to reconceptualise the key concepts of learning, retention and transfer. He adopts a phenomenological stance, suggesting that individual differences such as intelligence, prior knowledge and experience, attitudes and cognitive skills will have a crucial effect on what is learned and retained. The reason why true experiments come up with few substantive findings is that they employ procedures to randomise the very individual differences which determine what is learned and what is not. Beretta (1986) has made similar points in his call for the use of field rather than laboratory experimentation in language program evaluation.

Returning to the domain of language, rather than the more broadly conceived cognitive domain, the argument for program-free assessment is, to my mind rather curious. If the purpose of providing learners with a language education is to enable them to carry out a range of communicative tasks in that language, then it would seem entirely proper to base one's assessment on the achievement of specific curricular goals rather than on vaguely formulated notions of proficiency operationalised through proficiency scales and other tests of dubious validity. Such a suggestion is consonant with current trends in assessment outlined by Baumgart (1987):

> - a concerted move towards some form of standards-based assessment;
> - a growth in school-level initiatives in assessment and reporting, including quite widespread use of profiles, records of achievement and goal-based asessment;
> - much closer links between curricula and assessment with an emphasis on formative assessment;
>
> an emphasis on positive achievement and attempts to negotiate tasks and objectives which stretch students' capabilities but which also offer a reasonable chance of success;
> - consideration of the use of summative system-level records, albeit produced by schools, to underwrite and supplement formal certificates.
>
> (Cited in Brindley 1989: 93)

Brindley (1989) provides an invaluable source book of practical ideas, suggestions and illustrations of ways of incorporating criterion-related assessment instruments into the curriculum. He provides samples of performance profiles, records of achievement, graded objectives, rating scales, self-assessment checklists. Examples of such instruments from Nunan (1988) and Scarino et al. (1988) are provided in an appendix to the paper. Brindley himself has written extensively on the distinction between achievement testing and proficiency testing, arguing that the division fails to capture the range of purposes for which assessment may be carried out, and, further, that it fails to distinguish between the type and level of information. He attempts to resolve the tension between the two concepts by

postulating three different types of achievement / proficiency. Of these, only the first is "program-free". (Clark has coined the term "prochievement" to capture the idea of ongoing communicative assessment that is related to the program's proficiency goals.

> Level 1: Achievement of overall proficiency in a particular language skill or skills ("general" proficiency)
> Level 2: Achievement of particular proficiency-related objectives as part of a given course ("functional")
> Level 3: Achievement of specific objectives relating to knowledge and enabling skills taught in a particular course ("structural") (Brindley 1989).

Thus far, I have analysed and critiqued the notion of utilizing curriculum-free proficiency measures as means of assessing student progress. I have outlined some of the conceptual problems of the concept itself, as well as pointing out some of the inadequacies of instruments for measuring general language proficiency. It should be clear, therefore that I do not accept the validity of using such measures for the purposes of program evaluation. I would also refer you to Bachman's discussion on objectives-based and program-free evaluation. In the rest of the paper, I should like to focus more directly on program evaluation, and suggest that, while the incorporation of criterion-referenced assessment measures should form part of any adequate evaluation process, that they should not form the whole, or even the major part of the evaluation process. The two principal justifications I should like to offer for this assertion are (1) that evaluation involves much more than simply monitoring and measuring learning progress, and (2) that evaluation needs to focus on instructional processes as much as learning outcomes.

In concluding this section, I should like to point out that the use of individual gain scores to determine program effectiveness is not only problematic on theoretical grounds, but also on the practical grounds that gain scores are often not picked up due to the grossness of the measureing instruments. Within the Australian Adult Migrant Education Program, there are instances in which proficiency scores are actually lower at the end of a course than at the beginning!

The scope of program evaluation

In this paper, I have argued against a narrow input-output view of program evaluation, which references evaluation solely against learner output. The breadth and scope of any program evalaution must be referenced against two two important questions: "Who wants to know?" and "Why do they want to know?" As Cronbach has said, in his call for a reformulation and transformation in evaluation:

> The proper mission of evaluation is not to eliminate the fallibility of authority or to bolster its credibility. Rather, its mission is to facilitate a democratic, pluralistic process by enlightening all the participants. ... The evaluator is an educator; his success is to be judged by what others learn. .... Scientific quality is not the principal standard; an evaluation should aim to be comprehensible, correct and complete, and credible to partisans on all sides.
>
> (Cronbach 1980: 1, 11)

Assuming that most evaluations are not simply tokenistic exercises in indictment or exoneration, then program evaluators will want not only / even 'proof' in product terms, but 'insights' into the curicular processes and dynamics giving rise to particular outputs. In order to generate such insights, questions needs to be asked, and data gathered, on different aspects of the curriculum. Any area of the curriculum can be evaluated, from initial program planning through to the assessment / evaluation processes themselves. Some of the questions which might be posed in relation to different curriculum areas are set out in Table 1, which has been extracted from Nunan 1988.

Table 1
Some key questions in program evaluation

Curriculum area	Sample Questions
The Planning Process	
Needs Analysis	Are the needs analysis procedures effective? Do they provide useful information for course planning? Do they provide useful data on subjective and objective needs? Can the data be translated into content?
Content	Are goals and objectives derived from needs analysis? If not, from where are they derived? Are they appropriate for the specified groups of learners? Do the learners think the content is appropriate? Is the content appropriately graded? Does it take speech processing constraints into account?
Implementation	
Methodology	Are the materials, methods and activities consonant with the prespecified objectives? Do the learners think the materials, methods and activities are appropriate?
Resources	Are resources adequate / appropriate?
Teacher	Are the teacher's classroom management skills adequate?
Learners	Are the learning strategies of the students efficient? Do learners attend regularly? Do learners pay attention / apply themselves in class? Do learners practise their skills outside the classroom? Do the learners appear to be enjoying the course? Is the timing of the class and the type of learning arrangement suitable for the students?

	Do learners have personal problems which interfere with their learning?
Assessment and evaluation	Are the assessment procedures appropriate to the prespecified objectives? Are there opportunities for self-assessment by learners? If so, what? Are there opportunities for learners to evaluate aspects of the course such as learning materials, methodology, learning arrangement? Are there opportunities for self-evaluation by the teacher?

As I have already pointed out, in any evaluation, estimating the extent of learning outcomes is only a first step. Working out why certain learners have not achieved program goals is a much more difficult process requiring interpretation and analysis. In a study into teacher perceptions of the causes of learner failure reported in Nunan (1988), a group of ESL teachers were asked to nominate those causes which they felt were significant factors in the failure of learners to achieve program goals. The results of this investigation are summarised in Table 2. I have subcategorised these into causes attributable to the learner and causes attributable to the teacher.

Table 2

Survey results of causes of learner failure (After Nunan 1988)

Cause	Percentage of teachers rating this as a significant factor in learner failure
Causes attributable to the learner	
Inefficient learning strategies	77
Failure to use language out of class	77
Irregular attendance	45
Particular macroskill problems	32
Poor attention in class	9
Personal (non-language) problems	9
Learner attitude	4
Causes attributable to the teacher	
Inappropriate learning activities	32
Inappropriate objectives	27
Faulty teaching	23

From the data, it can be seen that, in general, the teachers surveyed saw responsibility for failure residing largely with the learners. (Although it is worth noting that, in relation to causes attributable to the teacher, one third of those surveyed identified inappropriate learning activities as a possible cause, and approximately a quarter identified inappropriate objectives and faulty teaching as having a significant effect on learning outcomes.)

The Need for Process Data in Program Evaluation

In order to validate the sorts of observations yielded by the study reported above, it is important to obtain data about learning and teaching processes themselves. Systematic observation is one important means of collecting such data. Non-observable problems such as failure to activate language out of class can be collected through learner diaries and self-reports. Other techniques, which are described and illustrated in some detail in Nunan (1989) include interviews and questionnaires, protocol analysis, transcript analysis, stimulated recall, and seating chart observation records. Ideally, a number of such techniques and instruments should be utilized in order to obtain multiple perspectives on the program under investigation.

The desirability of obtaining data on program outcomes and teaching processes is illustrated in a study reported in Spada (1990). This investigation sought to determine (a) how different teachers interpreted theories of communicative language teaching in terms of their classroom practice, and (b) whether different classroom practices had any effect on learning outcomes. Three teachers and their intermediate "communicatively-based" ESL classes were used in the study. Each class was observed for five hours a day, once a week, over a six-week period. Students were given a battery of pre- and post-tests including the Comprehensive English Language Test and the Michigan Test of English Language Proficiency. The study utilized the COLT observation scheme as well as a qualitative analysis of classroom activity types. This indicated that one of the classes, Class A, differed from the other two in a number of ways:

> A spent considerably more time on form-based activities (with explicit focus on grammar), while classes B and C spent more time on meaning-based activities (with focus on topics other than language). Classes B and C also had many more authentic activity types than class A. Furthermore, the classes differed in the way in which certain activities were carried out, particularly listening activities. For example, in classes B and C, the instructors tended to start each activity with a set of predictive exercises. These were usually followed by the teacher reading comprehension questions to prepare the students for the questions they were expected to listen for. The next step usually involved playing a tape-recorded passage and stopping the tape when necessary for clarification and repetition requests. In class A, however, the listening activities usually proceeded by giving students a list of comprehension questions to read silently; they could ask teachers for assistance if they had difficulty understanding any of them. A tape-recorded passage was then played in its entirety while students answered comprehension questions.
>
> (Spada 1990: 301)

The qualitative analysis confirmed the class differences, showing, for example, that class A spent twice as much time on form-based work than class C and triple the time spent by class B. To investigate whether these differences contributed differently to the learners L2 proficiency, pre- and post-treatment test scores were compared in an analysis of covariance. Among other things, results indicated that groups B and C improved their listening significantly more than group A, despite the fact that class A spent considerably more time in listening practice than the other classes.

Research such as that carried out by Spada indicated that there are in fact measurable differences in the way in which instruction is delivered in language programs which have similar ideological underpinnings, and that these differences can be related to learning outcomes. On a methodological level, it indicates that we need qualitative data based on classroom observation if we are to interpret, for the evaluative purposes of making decisions about program alternatives, the quantitative data yielded by assessment instruments of various sorts.

CONCLUSION

In this paper, I have taken a critical look at the role of second language proficiency assessment in program evaluation. I have examined some of the problematic aspects of the construct 'general language proficiency', as well as the theoretical and practical problems associated with attempting to measure such a construct. While I have referenced most of my comments against rating scales of one type or another, they are also pertinent to other types of proficiency test. As an alternative, I have suggested that curriculum-bound, criterion-referenced forms of assessment be developed. Sample assessment instruments are appended to the paper.

Given the length, purpose and nature of this paper, it has not been possible to comment on the problems associated with criterion-referenced assessment. I refer you to the paper given at this conference by Brindley who addresses some of the problems of trying to ensure validity and reliability. For example, how many times must a learner be observed to be able to do something, under what conditions, with what constraints, and in what contexts?

Assesment is an important component of program evaluation. However, determining what learners have or have not gained from a program is only one aspect of the evaluation process. In the paper, we have seen some of the o.her curricular elements which may fruitfully form the subject of any comprehensive evaluation.

In the final part of the paper, I argued that we need to collect information on teaching processes as well as learning outcomes. Techniques for collecting such data are outlined, and a study illustrating the importance of having both process and product data is reported. Ultimately, the type of evidence which is collected, and the ways in which it is interpreted and reported must proceed with reference to the purpose, scope and nature of the evaluation itself. If the principal purpose is to provide data to funding authorities for accountability purposes, the processes and outcomes are likely to be significantly different from an evaluation designed to provide feedback to teachers or one aimed at the development of new materials and teaching techniques.

REFERENCES

Bachman, L. 1989. The development and use of criterion-referenced tests of language ability in language program evaluation. In R. K. Johnson (ed.) The Second Language Curriculum. Cambridge: Cambridge University Press.

Baumgart, N. 1987. Emerging trends. In Reports and Records of Achievement for School Leavers. Project Newsletter No. 2, April 1987.

Beretta, A. 1986. A case for field-experimentation in program evaluation. Language Learning, 36, 3

Brindley, G. 1989. .Assessing Achievement in a Learner-Centred Curriculum. Sydney: National Centre for English Language Teaching and Research. CAL 1983. From the Classroom to the Workplace: Teaching ESL to Adults. Washington: Center for Applied Linguistics.

Chomsky, N. 1965. Aspects of the Theory of Syntax. Cambridge Mass.: M.I.T. Press.

Cronbach, L. 1980. Toward Reform of Program Evaluation. San Francisco: Josey-Bass.

De Corte, E. (ed.) 1987. Acquisition and transfer of knowledge and cognitive skills. International Journal of Educational Research, 11, 6.

De Cotre, E., H. Lodewijks, R. Parmentier and P Span (eds.) Learning and instruction. European research in an international context. Studia Pedagogica, 1.

Diller, K. 1978. The Language Teaching Controversy. Rowley Mass.: Newbury House.

Glaser, R. 1987. Learning theory and theories of knowledge. In E. De Corte et al. (eds.)

Gronlund, N. 1981. Measurement and Evaluation in Teaching. New York: Macmillan.

Higgs, T.V. (ed.) 1984. Planning for Proficiency: The Organising Principle. Lincolnwood: National Textbook Company.

Lantolf, J.P. and W. Frawley. 1988. Proficiency: understanding the construct. Studies in Second Language Acquisition, 10, 2.

Krashen S. 1981. Second Language Acquisition and Second Language Learning. Oxford: Pergamon.

Krashen, S. 1982. Principles and Practice is Second Language Acquisition. Oxford: Pergamon.

Nunan, D. 1988. The Learner-Centred Curriculum. Cambridge: Cambridge University Press.

Nunan, D. 1989. Understanding Language Classrooms. London: Prentice Hall.

Rea, P. 1985. Language Testing and the Communicative Language Teaching Curriculum. In

Y.P. Lee, C.Y.Y. Fook, R. Lord, and G. Low (eds.) New Directions in Language Testing. Oxford: Pergamon.

Resnick, L. 1987. Instruction and the cultivation of thinking. In E. De Corte et al. (eds.)

Richards, J. 1985. Planning for proficiency. Prospect, 1, 2.

Richards, J. C. and D. Nunan. (eds.) 1990. Second Language Teacher Education. Cambridge: Cambridge University Press.

Savignon, S. and M. Berns. (eds.) 1984. Initiatives in Communicative Language Teaching. Reading Mass.: Addison-Wesley.

Scarino, A., D. Vale, P. McKay and J. Clark. 1988. Evaluation, Curriculum Renewal and Teacher Development. Australian Language Levels Guidelines Book 4. Canberra: Curriculum Development Centre.

Spada, N. 1990. Observing classroom behaviours and learning outcomes. In J. C. Richards and D. Nunan. (eds.) 1990. Second Language Teacher Education. Cambridge: Cambridge University Press.

Thorndike, E. and R. Woodworth. 1981. The influence of improvement in one mental function upon the efficiency of other functions. Psychological Review, 8.

Voss, J. 1987. Learning and transfer in subject matter learning: A problem-solving model. International Journal of Educational Research, 11, 6.

Wolf, R. M. 1984. Evaluation in Education. New York: Praeger.

APPENDIX: Sample Criterion-Referenced Assessment Instruments

(Source: D. Nunan. 1988. The Learner-Centred Curriculum. Cambridge: Cambridge University Press.)

TABLE 9.1

Sample rating scales

Indicate the degree to which learners contribute to small-group discussions or conversation classes by circling the appropriate number.

(Key: 5 – outstanding, 4 – above average, 3 – average, 2 – below average, 1 – unsatisfactory)

1	The learner participates in discussions.	1	2	3	4	5
2	The learner uses appropriate non-verbal signals.	1	2	3	4	5
3	The learner's contributions are relevant.	1	2	3	4	5
4	The learner is able to negotiate meaning.	1	2	3	4	5
5	The learner is able to convey factual information.	1	2	3	4	5
6	The learner can give personal opinions.	1	2	3	4	5
7	The learner can invite contributions from others.	1	2	3	4	5
8	The learner can agree/disagree appropriately.	1	2	3	4	5
9	The learner can change the topic appropriately.	1	2	3	4	5

Rate the learner's speaking ability by circling the appropriate number.

1 2 3 4 5 6 7 8 9 10

Incapable of carrying out simple conversation

Carries out simple conversation giving personal information

Rate the learner's listening ability by circling the appropriate number.

1 2 3 4 5 6 7 8 9 10

Incapable of following simple instructions

Follows simple instructions in classroom setting

Checklist of reading skills

YES	NO	Recognises Roman script upper/lower case
YES	NO	Identifies numbers in various formats
YES	NO	Comprehends key content words/phrases in context
YES	NO	Retrieves simple factual information from short texts
YES	NO	Comprehends regular sound/symbol relationships
YES	NO	Sight reads key function words
YES	NO	Identifies genre of common texts
YES	NO	Identifies topic of simple text on familiar subject
YES	NO	Uses alphabetical indexes
YES	NO	Follows written instructions

(Source: Scarino, A. et al. 1988. Evaluation, Curriculum Renewal and Teacher Development. Australian Language Levels Guidelines Book 4. Canberra: Curriculum Development Centre.)

Table 16: Performance indicators

Content		
• Completion of activity	activity not completed	activity totally completed
Quality of performance		
Communication goals		
• comprehension of information (from interlocutor or text)	minimal comprehension	total comprehension
• intelligibility of response	minimally intelligible	totally intelligible
• quality of language resource:		
degree of accuracy (including grammar, vocabulary, pronunciation)	minimal accuracy	high accuracy
degree of fluency (speed and rate of utterance, ability to structure discourse)	minimal fluency	high fluency
range of expression (ability to go beyond stereotyped forms and to generate language)	limited range	good range
Sociocultural goals		
• sociocultural appropriateness	inappropriate	appropriate
• sociocultural knowledge	minimal knowledge	good knowledge
Learning-how-to-learn goals (including skills and strategies)		
• use of communication strategies	minimal use	effective use
• level of support required	strong reliance on support	no support required
General knowledge goals		
• knowledge of subject matter of the activity	minimal knowledge	good knowledge

(Source: Scarino, A. et al. 1988. Evaluation, Curriculum Renewal and Teacher Development. Australian Language Levels Guidelines Book 4. Canberra: Curriculum Development Centre.)

Table 18: General criteria for judging performance in activity-type 2

Activity-type 2	General criteria
Participate in social interaction related to solving a problem, making arrangements, making decisions with others, transacting to obtain goods, services, and public information (interacting and deciding)	Conversation activities • Did the learner succeed in solving the problem/making arrangements/ arriving at a decision/obtaining the particular goods or services? • Did the learner understand the information provided by others? • Were the learner's utterances intelligible? • Were the learner's utterances sufficiently accurate so as not to interfere with conveying meaning? • Were the learner's utterances appropriate to the sociocultural context? • Did the learner's responses cohere with the flow of the discussion? • Was the learner able to interact with others, take turns, maintain the conversation, generate questions, build on ideas? • Did the learner need help from others? • Did the learner provide information for the discussion?
	Correspondence activities • Did the learner complete the activity set? • Did the learner understand the information provided in the stimulus? • Was the learner's response intelligible? • Was the learner's response sufficiently accurate so as not to interfere with conveying meaning? • Was the learner's response appropriate to the sociocultural context? • Was the learner's response coherent? • Did the learner need support from the stimulus model or dictionary (if provided)?

Table 19: General criteria for judging performance in activity-type 3(a) & 3(b)

Activity-types 3a & 3b	General criteria
3a Obtain information by searching for specific details in a spoken or written text, and then process and use the information obtained (searching and doing) 3b Obtain information by listening to or reading a spoken or written text as a whole, and then process and use the information obtained (receiving and doing)	• Did the learner understand and extract the relevant information relating to the activity set? • Did the learner reproduce the information, as required by the activity? • Did the learner make an appropriate decision/choice/response on the basis of the information obtained • Was the learner's response intelligible? • Was the learner's response sufficiently accurate so as not to interfere with meaning? • Was the learner's response appropriate to the sociocultural context? • Was the learner's response coherent? • To what extent did the learner need support from others (interlocutor, or spoken or written text)?

Note: all macroskills are implied in these activity-types. Responses may be oral or written.

HOW PROGRAM PERSONNEL CAN HELP MAXIMIZE THE UTILITY OF LANGUAGE PROGRAM EVALUATIONS

Ronald Mackay

1. THE SOURCE OF THE CONTENT OF THIS PAPER

This paper is based on my own experience as a researcher and as an evaluator in the field of language teaching and applied linguistics. It has been derived directly from a textbook which I am in the process of preparing for publication. My interest in language program evaluation has been a natural development arising out of my work in first teaching, then designing curricula and in materials writing. Over the past decade I have been involved in a dozen program evaluations. Some of these have been small-scale, involving individual programs, small numbers of teachers and students, offering minimal financial resources for the evaluation activity and only one evaluator. Other evaluations have been quite extensive, involving entire provincial and territorial school systems, scores of schools, and budgets in the hundreds of thousands of dollars.

2. THE PURPOSE OF THIS PAPER

The purpose of this paper is **not** to outline to you what you must learn in order to become a program evaluator yourself. There are dozens of "How To .." books in the field of programme evaluation. Some are good, some bad and some merely indifferent.Furthermore, not everyone wants, and still fewer have the opportunity to leave teaching or program management for the full-time practice of evaluation. On the contrary, this paper is directed at those who staff language programs which are likely to be evaluated at some time or other. It is directed at teachers, program planners, materials writers and program managers in order to help them understand what might happen when their program comes up for evaluation. Its purpose is to help them overcome their understandable concerns and fears urrounding the evaluation process; to show them ways in which they can cooperate with the evaluator and also ways they can defend themselves against inadequately trained or insensitive evaluators. Most of all, its purpose is to provide them with specific practical and workable suggestions as to how they can become active participants in any evaluation of their program. Participation helps to ensure that the results of the evaluation will address their concerns, their questions and their interests, reflect their perceptions and contribute to their aspirations, and not merely those of some outside evaluator who, like the proverbial cat, crept in, evaluated, and crept out again.

3. SOME PRELIMINARY QUESTIONS

3.1 Why Do I Consider It Necessary For Program Staff to Learn What I Have, Only Semi-Humourously, Called "Self-Defence" ?

First of all because program evaluation in our field of second language teaching and applied linguistics so far offers little in the way of training for potential evaluators, and has articulated no professional standards with which evaluators are uged to comply. Hence many different kinds of person can assume the title of 'evaluator' and many practices of varying kinds and quality are undertaken in the name of 'evaluation'.

Rather than focus upon the matter in a negative way, I will briefly outline to you the posiitive steps that are being taken in Canada to deal with the growing interest in evaluation. Six years ago the Canadian Evaluation Society was established with these objectives :

* To provide a forum whereby program managers, administrators, policy-makers, practitioners teachers and students can discuss the theory and practice of evaluation in Canada;

* To promote the high quality of evaluation of public and private programs throughout the country;

* To develop theories, standards, and practices of evaluation;

* To promote training programs in planning the design, strategy, methods, anàlysis and application of results for all types of evaluation;

* To provide a forum of exchange on policies, practices, applications and sources of funding for evaluation.

The C.E.S. draws its membership from a diverse group of professionals in the public private and academic communities who share a common interest in evaluation They represent disciplines ranging from psychology sociology, social work, economics, health sciences, administration, political science, and policy sciences to accounting, engineering and urban and regional planning. The C.E.S. holds a national convention, publishes a journal and a newsletter, and is organized around regional groups which arrange meetings and professional development sessions for their members.

This is one way of tackling the growth of and interest in evaluation - namely to encourage the professional development of those who practice evaluation. This paper is another, complimentary way of tackling the issue - namely it encourages the interest of program staff in becoming more informed about evaluation and its implications so that they can participate with the evaluator in the evaluation process and so that evaluation is not something that is done to them, but for them and with their consent, cooperation, and understanding.

3.2 "What is the Difference Between Research and Evaluation ?"

Applied linguistics research and second language program evaluation are for the most part two different and distinct activities.They can be distinguished from eachother on a number of counts. Some of the differences listed below (see also, Popham 1975; Patton 1986; Worthen and Sanders 1987) overlap somewhat, and the list is not exhaustive.:

THE IDENTIFIABLITY OF THE CLIENTS

Research is not usually carried out for an identifiable client. Research tends to be funded by governmental or philanthropic organizations, but it is not carried out for their benefit. An evaluation is funded by a sponsor who may be equally as faceless as a government department or a philanthropic organization but is carried out for a specific client or group of clients, one of whom may or may not be the funder. These clients are identifiable as individual people with a particular identifiable interest or stake (as it is called) in the results of the evaluation.

THE RELATIVE INDEPENDENCE OF THE INQUIRER

Researchers in effect work for themselves in the sense that they decide what question to research, what focus they will adopt, what variables they will manipulate, which they will hold constant and which they will ignore altogether. Evaluators on the other hand answer questions posed by stakeholders, focus on the latters' conserns and interests and examine issues selected by those funding, managing, teaching in or in some other way affected by the program. Researchers normally initiate their own research projects based on their own personal curiosity and interests. They decide what they want to investigate and then seek the funds from an appropriately sympathetic organization to carry out the investigation. Evaluators respond to the information needs of others, normally stakeholders in a particular program.

THE USEFULNESS OF THE INQUIRY

Research is not motivated by immediate utility, whereas evaluation is. The results of an evaluation will be used by a client or by different groups of clients (called the stakeholders) to answer their questions, confirm their suspicions, and inform their decisions concerning one or more aspects of the program that they design, manage, operate, teach in, learn in or are in some way or another affected by.

THE MOTIVATION OF THE INQUIRER

The researcher is motivated by a desire to expand the frontiers of human knowledge; the evaluator is motivated by the desire to provide illuminating answers to specific questions posed by specific stalkeholders about a specific program in a particular context.

THE OBJECT OF THE INQUIRY

Research strives to further knowledge, serve truth and broaden our understanding of the world and the types of phenomena (e.g. types of programs) and so tries to focus on the typical obtained by means of random sampling (or some similarly scientific strategy of selection) within the populations they wish to generaluize to. Evaluation, on the other hand, is concerned with the interests and concerns of specific people involved with specific programs and so focuses on the individual project within its own unique context.

GENERALIZABILITY

The researcher seeks to uncover the general laws of nature usually by means of establishing significant relationships between variables. Evlauators focus on specific contextrs, particular programs with all their individual constraints and idiosyncracies. Whereas the researcher wants to make strong and broad generalizations the evaluator wants to enlighten individual stakeholders involved with individual programs. Researchers seek to uncover general laws which govern human learning. Evaluators seek to describe a specific program identifiable in time and place, in an illuminating way. and, in the process, to provide stakeholders with a better understanding of its strengths and weaknesses.

THE ROLE OF EXPLANANTION

Researchers seek to explain the stable relationships between variables ideally in a cause and effect relationship. Evaluators seek to provide credible exlanations to stakeholders for what is occurring in their program so that specific adjustments might be made to improve it.

RELEVANCE OF TIME"

More often than not, the researcher works to self-imposed deadlines. These deadlines derive from factors extrinsic to the research - dates related to the university calendar, dates related to the annual calendar for funding by grant agencies.The value of research results are unlikely to be wiped out if the research project is completed a little behind schedule. Evaluators work to deadlines intrinsic to the project that they are evaluating. Stakeholders require information by certain times so that effective decisions can be made. Information received from a tardy evaluator one day after a major project planning meeting of the stakeholders may render that information generated by the evaluation totally useless.

DISCIPLINE LOYALTY"

Research normally is undertaken within the well-defined (but often arbitrary) paremeters of specific disciplines. Evaluations seldom demand loyalty to one particular discipline and often involve the evaluation team working cooperatively across specialization boundaries.

SOME REASONS FOR THE CONFUSION BETWEEN EVALUATION AND RESEARCH

First, training courses for second language program evaluators are uncommon but instruction in research methods based on the experimental paradigm are not. Thus even the m ost willing of evaluators may be the victims of one particular, and very limited, frame of reference. Second, when research funds are tight, enterprizing researchers may seek evaluation contracts to continue their research work, consciously or unconsciously defining the focus of the evaluation to coincided with their own interests. Third, some sponsors of evaluations may make money available for an evaluation without providing adequate terms of reference for the evaluation, thus leaving the researcher to determine the focus and direction of the activity him or herself.

3.3 What is the difference betweenm functioning as an evaluator and functioning as an advocate?

Evaluation and advocacy are legitimate activities and both may be conducted in a professional manner. The function of an advocate however is to present a program in the best possible light in order to influence the decisions made by a major information user - usually the funder - in a particular way - usually to continue funding or to increse the funds for the program in question. The function of an evaluator cannot be so one-sided. An evaluator who functions as an advocate will lose his/her credibility as an independent judge.

Some professionals, those who are committed to one of the many forms of social reform, for example, may believe that their moral duty is to advocate. There is nothing wrong with that so long as they are up-front about their position and do not present themselves or their activities as evaluation.

4. A RESEARCH-LIKE PROCEDURE WHICH MASQUERADES AS EVALUATION

When research masquerades as evaluation, it tends to follow a particular process (Fig. 1).

FIGURE 1

RESEARCH - LIKE EVALUATION PROCEDURE

FUNDER INITIATES

EVALUATOR DECIDES FOCUS

EVALUATOR DESIGNS EVALUATION

EVALUATOR GATHERS DATA

EVALUATOR ANALYSES DATA

EVALUATOR INTERPRETS DATA

EVALUATOR MAKES RECOMMENDATIONS

EVALUATOR WRITES REPORT

EVALUATOR DEPOSITS REPORT

FUNDER RECEIVES REPORT

RESEARCH-LIKE EVALUATION PROCEDURE

This not uncommon research-like procedure which is sometimes passed of by its practitioners (purposely or unconsciously) as evaluation has little time for the stakeholder's concerns and allows little opportunity for input from the program staff except at those points where the researcher unilaterally decides that it is necessary for his or her purposes to obtain information, usually of a predetermined kind, from them.

5. How can program staff make a difference ?

Thus far we have painted a pretty dismal picture - research which masquerades as evaluation, evaluators who are insensitive to the stakeholders, advocates in disguise, funders who disburse funds without due care and attention to what will be given in return....... How can we as program staff change the picture ? How can we as program staff ensure that the researcher does not run away with the evaluation funds for his or her own purposes which may have marginal or even no relevance to the purposes of the principal infotrmation users and the other stakeholders ? How can we as program staff get what we want and need out of a program evaluation ?

The remainder of this paper will be devoted to suggesting ways that program staff can go about gaining a foothold in the evaluation activities and by so doing, become instrumental modifying, to a greater or lesser extent, the **research-like process.** I will deal with ways staff can intervene in two parts. The first part will deal with early and successful intervention and will show the **cooperative evlauation process** which results. The second part will deal with later and less successful intervention, and will illustrate what I call the **utility enhanced evaluation process.** In this paper I will not deal with a third part which shows program staff what action might be appropriate when it becomes clear that their attempts at gaining access to the evaluation process are doomed to failure. You will have to buy the book to find that out !

5.1 Turning the Research-Like Procedure Into A Cooperative Process

I will take the research-like process discussed above and suggest how, early on in that process, the stakeholders can gain a foothold and so help ensure that the evaluation results in a report that can be used by them.

Let's start at the very beginning at the point where the **Funder Initiates an Evaluation**, in Fig. 1.

For any program, somebody holds the purse strings to the fund from which any evaluation to be undertaken will be financed. Sometimes the holder of the purse strings is not a major information user or even not a stakeholder at all. This is the case in many government sponsored programs. In yet other cases, the funder is a stakeholder and a potential information user and may know virtually nothting about evaluation and so hires an evaluator to take control. Sometimes the funder simply assumes that he or she is the **only** person interested in the evaluation and does not even consider that others may have a substantial stake in any inqiry and may have questions which they would like answered in order to make program -related decisions.

This stage, because it is the point at which the evaluation is initiated, is the most crucial stage at which program staff can help to influence the entire subsequent evaluation process.The very first task of the enterprising and concerned stakeholder is to find answers to the following questions :

#1 FIND ANSWERS TO THE FOLLOWING QUESTIONS :

Is it planned that the program will be evaluated ?
If so, when is it scheduled to happen ?
Where will the funds come from ?
How much will they amount to ?
Are there any constraints upon the type of evaluation that must be carried out ?

For example does the plan specify whether it will be an internal or an external evaluation ? Will it be formative or summative ? Will it be undertaken for a predetermined purpose, e.g. to determine whether the program should be expaned to other schools, provinces or states ? or whether there are more cost-effective ways of achieving the same results ? or whether the program is achieving the purposes for which it was originally mounted ?

When you have found accurate answers to these questions, and you have established that there is the potential for your interests in the program to be addressed,ask yourself, " What do I need to know about this program ?" and "How will I use that information ?" Notice that the operative words here are **'need'** and 'use'. It is as a potential **user** of evaluative information that you have the right to intervene or to gain access to the evaluation process, not simply as a curious by-stander. For example, as a teacher you might need to know if the strategy of selective error-correction that you are employing is having immediate and or lasting effects on student production so that you can continue, modify or discontinue the practice; as a materials writer or as a head teacher, you might need to know if your materials are being used in the way and under the conditions they were intended to be used, etc..Once you have identified the information you need :

#2 LET THE FUNDER KNOW THAT YOU HAVE SPECIFIC QUESTIONS YOU REQUIRE ANSWERS TO, OR ISSUES THAT YOU WISH TO HAVE ADDRESSED.

In other words, let the funder know that you are a stakeholder, and precisely what your stake is in the evaluation This may come as a surprise to a funder who has not given consideration to any program staff as being stakeholders.

#3 TELL THE FUNDER HOW YOU WOULD MAKE USE OF THE INFORMATION GENERATED BY THE EVALUATION

In other words, let the funder know that you have given serious consideration to your stake in the evaluation and that addressing your concerns or questions is important for the future of the program.

If, at this stage, the funder embraces you as a full partner in the evaluation process, most of your serious problems are over. You have switched right over from the danger that a **research-like procedure** will be employed to the liklihood that a **cooperative procedure** **(Fig. 2)** will be adopted.

Figure 2

COOPERATIVE EVALUATION PROCEDURE

FUNDER INITIATES

CLIENT AND STAKEHOLDERS IDENTIFY PIU'S

CLIENT AND PIU'S IDENTIFY Questions and Concerns

CLIENT AND PIU'S IDENTIFY USE

CLIENT AND PIU'S DRAW UP Request for Proposals

CONTRACTORS SUBMIT TENDERS

CLIENT (AND PIU'S) SELECT WINNER

CLIENT DRAWS UP CONTRACT

CLIENT AWARDS CONTRACT

EVALUATOR CONDUCTS EVALUATION ACCORDING TO CONTRACT

But do not relax your attention or your guard! There is a great deal of solid work to do, however. Once the funder has agreed that the concerns other than simply those of the funder should be addressed, the matter of who the other principal information users (PIU's) are and the issue of who addresses their questions (i.e. who the evaluator is), is no longer a trivial question. Any evaluator hired, has to be able to address the stakeholders' concerns and questions, and tackle the task of answering them in a manner appropriate to your requirements, within the time constraints imposed by the need for subsequent action, and within the financial constraints imposed by the evaluation budget. No small task to administer [1]

The formal way of finding an evaluator to do all this, is by listing the requirements of the stakeholders into a document called a Request for Proposals usually referred to as the RFP. The minimum contents of the RFP are :

1. a description of the context of the program

2. a clear statement of the purposes of the evaluation

3. a list of the stakeholders in the evaluation

4. a list of the principal information users

5. the time frame within which the evaluation must be completed

6. the financial resources available for the evaluation

7. the form(s) in which the final report(s) is/are to be presented.

Potential evaluators are then invited to respond, at their own expense, to the RFP in a competitive process known as 'tendering'. The tender is usually advertised in periodicals read by the kind of professional you are seeking to attract. Alternatively, the stakeholders might suggest three or four names of individuals or groups known to have the skills required to address the contents of the RFP.

Adequate responses to the RFP should :

1. be clear, concise and jargon-free

2. clearly itemize all the costs involved in the proposed evaluation

3. show the qualifications and experience of the evaluator(s) to undertake the evaluation

4. clearly specify the tasks that would be undertaken (and their sequence) in order to complete the evaluation

5. give time-lines for each task

6. describe the methodology that would be used and how it would achieve the purposes of the evaluation

7. demonstrate convincingly that all the questions in the RFP would be answered appropriately

You and the funder, in your new cooperative relationship, can examine the potential evaluators' responses to the RFP and choose the one that meets your needs most closely. Once an evaluator has been chosen, the funder enters into a formal contractual agreement with him/her. The contract will, as a minimum. specify :

1. what the duties of the evaluator are

2. what the responsibilities of the funder are

3. what the total budget will be

4. how and when the evaluator will be paid

5. how the evaluation work plan can be amended (if at all)

6. who the person acting as liason between the evaluator and the funder is

5.2 Turning the Research-Like Procedure into a Utility-Enhanced Procedure

Let us imagine, for the purposes of the rest of the paper, that the funder has retained the services of an evaluator without consulting the stakeholders. What then ? How can you, the stakeholders gain a foothold in the evaluation process ? For the purposes of this paper, I will not deal with every step of attempting to convert the Research-like process into a utility-enhanced process, nor will I give a large number of suggestions or examples. I am sufficiently realistic to realize that if I were to be exhaustive, you would go to sleep, and sufficiently mercenary to fear that you might not buy my book when it appears !! So I will be relatively succint, and selective.

So let us start one step on in the research-like process where **The Funder Retains the Services of an Evaluator (Fig 1.)**

At this point it is invaluable to have the stakeholders participate as a group, identify their needs and concerns, become familiar with the needs and concerns of their fellows, and agree, if at all possible, upon a hierarchically ordered of questions that they would like to see addressed. So,

#4 MEET WITH THE OTHER STAKEHOLDERS AND HAVE THEM IDENTIFY WHAT INFORMATION THEY REQUIRE FROM THE EVALUATION AND HOW THEY WILL USE THAT INFORMATION

It must be borne in mind that the stakeholders **must** be able to show how the answers to their questions will be used to make decisions about the program. Questions asked out of curiosity, no matter how interesting they may be, or information which cannot be used are not legitimate concerns to present to the evaluator. For example, the project manager might like to know whether the teachers with more experience, or the teachers with more recent qualifications are having the greater success in implementing a new set of materials involving non-traditional classroom interaction. However, if because of union agreements there is no possibility of removing the less successful group or even of insisting on obligatory in-service training for them, the infomation, interesting though it may be, is not capable of being acted upn and so the question is not a legitimate one for the evaluation to address.

#5 TALK TO THE EVALUATOR

Program staff may feel that it is somehow "not fitting" to talk to the evlaluator about the program and about his/her plans for evaluating it; or they may feel discouraged by the evaluator appearing to weant to hold them at arms length from his/her activities. However, not only is it fitting to talk to the evaluator, it is essential to talk to him or her.

#6 ASK THE EVALUATOR TO EXPLAIN TO YOU AND THE OTHER STAKEHOLDERS HIS/HER APPROACH TO AND VIEWS ON 'GOOD EVALUATION'

A good evaluator can explain where he/she is coming from and what he/she is doing, **in the language of the stakeholders**. If the evaluator cannot, then the evaluator is not worth hiring. An evaluator who has one and only one perspective on evaluation and tries to impose that upon the program and the program staff is not worth hiring. But (for the purposes of this paper at least) the evaluator has already been retained ! So :

#7 LET THE EVALUATOR KNOW THAT YOU HAVE SPECIFIC QUESTIONS TO WHICH YOU REQUIRE ANSWERS OR CONCERNS WHICH YOU WOULD LIKE TO SEE ADDRESSED.

#8 LET THE EVALUATOR KNOW HOW YOU WOULD MAKE USE OF THE INFORMATION GENERATED BY THE EVALUATION IF IT WERE TO ADDRESS THESE QUESTIONS AND CONCERNS

A good evaluator will be interested in the perspectives of the stakeholders - their views on what they believe evaluation to be, how it is carried out, how they conceive of their program, what their concerns are and what they believe their infomation needs to be and why.If you let the evaluator know that you are counting on the results of the evaluation to obtain specific information which is otherwise unavailable in order to take action in some area of the program it is likely that he/she will not ignore your request.No evaluator worth his or her salt will choose to ignore questions asked by stakeholders who are also principal information users in favour of self-generated questions. A good evaluator will help the stakeholders to discover the right relationship between his/her expertise and their own contribution. A good evaluator will appreciate the feeling of self worth that can be fostered by encouraging the stakeholders to participate in the evaluation. Conversely, he/she will understand the feelings of suspicion, hostility, and disempowerment which will be engendered in the stakeholders if they are held at arms length, their concerns spurned, their questions ignored and their participation rejected.One of the principal advantages to be gained from an evaluation is the enlightenment gained by the stakeholders from an understanding of and participation in the process itself.

If, at this point, the evaluator demonstrates a willingness to work with the stakeholder group as partners in the evaluation, you have achieved an early entry into the **value-enhanced process.** It is clearly by no means an identical process to that described as cooperative above (Fig. 2), because you are stuck with an evaluator you did not select on the basis of a request for proposals and this evaluator despite his or her current willingness, may never have worked cooperatively with a stakeholder group before, or he/she may have a very limited perceptual framework for conducting evaluations, or he/she may at any time feel threatened by the new relationship and wish to end it. Time prevents me from offering you suggestions as to how to handle the various problems that can arise from this point on. Time also prevents me from offering you suggestions as to what you can do if your appeals to the evaluator at this step are unsuccessful. Suffice it to say that what we have called the value enhanced model of programme evaluation is limited in its usefulness to programme staff in proportion to the point at which you gain entry to the process. The earlier you gain entry, the greater the opportunity for you to influence the focus of the evaluation, the type of data collected and to participate in the interpretation phase and therefore the more potential for utility the results of the evaluation will have.

6. CONCLUSION

You will, however, no doubt appreciate that if you have failed at this point, it may be even more difficult (but certainly not impossible) to gain access to the process at one of the later stages. Nevertheless, the potential for increasing the utility of an evaluation diminishes in direct relation to the tardiness of the phase at which you get entry. It is likely that there is a point of diminishing returns and a whole new strategy involving a different type of thinking is required if you should be so unfortunate as to reach that point without having persuaded the evaluator to allow you into the process. If you are unsuccessful in gaining access to the evaluation process at all, then you simply end up with a researcher-directed study masquerading as evaluation! Such a study may be instrumental in furthering the researchers' interests by providing him/her with a publication or even a higher degree, but it will be unlikely to be of use or even of much interest to programme staff concerned about improving their activities or finding answers to questions which they require to make informed decisions.

I hope, however, that I have been successful not only in persuading you as program personnel that it is worth gaining access to the evaluation process, but also in showing you some practical ways by means of which you might successfuly become a respected partner in the evaluation of your own program.

REFERENCES

Patton, Michael Quinn. 1986. Utilization-Focused Evaluation. 2nd Edition, Beverly Hills, CA; Sage

Popham, W. James. 1975. Educational Evaluation. New Jersey; Prentice-Hall

Worthen, Blaine R. and Sanders James R. 1987. Educational Evaluation. N.Y. Longman

THE DEVELOPMENT OF SELF ASSESSMENT SKILLS in TESOL Teacher Preparation

Alastair L McGregor

INTRODUCTION

I believe that most of us involved in the field of teacher education and development would admit to a considerable degree of frustration and disappointment about the rather meagre outcome of all our efforts, the students involved often being the most frustrated and disappointed of all. Our experience (and reading of the thoughts and experiences of others) are wide, our planning, and usually execution, are thorough and vigorous, and our ideals are undoubtedly high. As the outcome of the process we hope to see competent teachers with a thorough command of content and teaching approaches, and with the analytical skills which would enable them to choose and execute these appropriately in any particular situation, and with these we long to observe the development of a high degree of self awareness, initiative and true sensitivity to the needs and personalities of their students. We like to think of ourselves as being at the cutting edge of change and development not only in our individual students but in the profession. (I am speaking broadly about both pre-service and in-service or continuing teacher education).

Yet, as I say, many of us - while too experienced and sensible to expect total success - nevertheless experience real anxiety at what seems to us to be an unacceptably high level of failure to achieve these objectives with any but a few of our students. We have a sense of casting our pearls before rather unappreciative swine, and grimly hope they are not synthetic pearls. I will not defend my assertion that this is a common feeling throughout the ranks of teacher educators (though I believe much writing confirms it) but at any rate admit to you that after 25 years in this section of the profession it is certainly my observation about my own courses and students.

My colleagues and I have therefore set about a modest reappraisal of our principles and approaches, one outcome being the procedure I should like to describe to you a little later, a procedure intended to build greater skills of self observation, self-analysis and self assessment, in our students.

It might be more important, however, to speak first of the particular strands of thinking and research which have influenced our re-evaluation. No sudden lights from heaven have struck us, of course, but over the past year or two we have become more aware of several elements that, even if not entirely absent or totally ignored by us in the past, have probably not been given enough weight in our thinking and procedures.

Firstly there is the increasing disillusionment with the series of approaches and methodologies that seem to pursue each other across our horizon with almost monotonous regularity, like sunshine and shadow on a windswept plain (as Conrad says). I hasten to say that this is not because we have not found these approaches useful in themselves. Far from it. Rather disillusionment has followed the never failing claim, or at least implication, that the latest approach is 'the' answer. It is the exclusivity of each approach that has come to raise a weary smile on the face of us practitioners. We have slowly learned that the communicative approach is not to be 'the' answer (For 'communicative' read functional/notional, cognitive code, direct etc etc according to your taste.) We have long ago learned (what classroom practitioners have always recognised) that all these approaches come to us value-laden and are in many cases totally unsuitable without major modification for application in the vast variety of socio-cultural situations in which they are supposed to operate. One thinks for example, of the stimulating (if slightly exaggerated!) analysis by Dr Sampson at a previous seminar of the values implicit in the communicative approach. (Sampson 1984). Yet this attitude of exclusivity persists. Some of us are old enough to remember the pronouncements of the high priests or gurus of the audio-lingual approach when teachers dared to suggest it wasn't (to put it mildly) being altogether successful! This was heresy. We know it's the answer, we've proved it not only on a theoretical base of learning theory and linguistic theory, but in our practical success with thousands of service men learning languages this way to meet post-war (or post invasion) needs. You must be doing it wrong! Remember? - some of you?

Nor, in spite of decades of experience has that kind of attitude been finally laid to rest; it just gets re-attached to the latest orthodoxy.

So, a few weeks ago I received this written comment from an external examiner (who shall remain nameless) on some examination candidates

> "..........candidates itemised anticipated language use and language to be modelled by the pupils from the teacher. Such a view of ESL teaching completely denies (!) the role of peer group interaction and the learning that pupils do outside the classroom, from the community, media etc. It represents a return to (a named approach) which was shown to be (!) restrictive and based on erroneous notions of the language learning process."

Nor is the disillusionment merely impressionistic and ill-founded. A never-ending stream of research, hundreds of thousands of man-hours of effort and perspiration have gone into the attempt to show the superior effects of some particular approach. Perhaps the best summary of the situation is to be found in Allwright's (1988) typically incisive account of the history of observational studies in the classroom and the very common failure (to the disappointment of those who wanted to prove otherwise) to establish any particular superiority in terms of language learning for one or the other approach. So obvious was this that in surveying a large number of studies, Long (1983) even asked the question "Does Second Language Instruction Make a Difference?" While the answer would, fortunately for us teachers, seem clearly to be yes, there was no clear evidence that any one particular form of instruction made more of a difference than another. Thus the days of confidently presenting, illustrating, modelling, encouraging in practice AN approach for our students have gone. Instead we find ourselves with a bank of approaches and activities which all have their usefulness when applied to appropriate situations and needs. We all, I trust, remain eternally grateful for Chomsky's memorable warning 25 years ago

> "In general the willingness to rely on 'experts' is a frightening aspect of contemporary political and social life. Teachers, in particular, have a responsibility to make sure that ideas and proposals are evaluated on their merits, and not passively accepted on grounds of authority, real or presumed.
>
> ......There is very little in psychology or linguistics (and dare we add, methodology and curriculum) that he can accept on faith" (Chomsky, 1966). (author's addition)

A second important influence has been the increasing emphasis that we ignore the learner's contribution at our peril, and I speak of the learner in the individual and group sense. The learners' needs, perceptions, moods, learning styles and strategies, as we well know, can very quickly make a nonsense of our thoroughly prepared curricula and lesson plans. Here, of course, is what makes us look so foolish when we try to insist on a method, an approach, indeed even when we are trying to establish whether any language teaching behaviour or device could be classified as 'good' or 'bad'! Politzer (1970) tried to do that and was forced to the conclusion that there were few, if any, absolutes:

> "In other words the very high complexity of the teaching process makes it very difficult to talk in absolute terms about 'bad' and 'good' teaching devices...
> The 'good' teacher is one who can make the right judgment as to what teaching device is the most valuable at any given moment."

or as Allwright (1972) put it when discussing the effects of using a particular technique or method:

> "It is, however, clear that much more than this is happening. People are interacting in a multiplicity of complex ways, as people, getting bored or even excited, getting encouraged or discouraged, more confident or less confident, and so on. It is a commonplace to assume that such events are important to learning, probably crucial, but this seems to have been largely left out of research on methodological comparisons."

It has fallen to me several times at preceding seminars here to emphasise this aspect from our own experiences in Australia and elsewhere and my observation is that people on the whole listen politely then shrug it off as too complex a matter to take into account seriously. I suggest to you that we dare not fail to take it into account no matter what degree of complexity it introduces. Let me give you a small group example from just a few weeks ago - Several of my colleagues and I take some advanced English classes for fairly high level, mostly professional migrants to Australia. It fell to one of my colleagues, (a new colleague, by the way,) to take a unit on the Arts, the language of the Arts. (The class was composed of a mixture of Eastern Europeans, Chinese, Indonesians, students from the Middle East and so on). He is an open and progressive teacher so on his first visit to the group he invited them to suggest what topics they would like, felt they needed, to have covered. Before the afternoon was finished a delegation from the class arrived at the door of the Director of Programmes. "Please remove this teacher; he doesn't know what he should be doing. Wants US to tell HIM... etc." We experienced this on a larger scale in Australia when a complete and fine curriculum prepared for the Adult Migrant Education Service more or less had to be abandoned. It had been prepared mostly with Vietnamese and Cambodian refugees in mind. Suddenly a great number of Eastern European refugees started arriving. They said "We don't want to know about supermarkets, and visits to doctors' surgeries. Just tell us - Is this the subjunctive or not?!" We cannot ignore the learners, plan in a vacuum or even plan on our perceptions of their needs. They may have different ideas; they march to the beat of another drum altogether, from the one we are banging so enthusiastically: they learn what they want to learn, not what we want them to learn. Remember the findings of Felix who concluded that much of the linguistic output of school language learners "could only be understood as essentially random behaviour, but that otherwise their classroom use of language suggested that they were using 'natural' processes of language acquisition rather than those the teaching was designed to promote." (Felix, 1981).

A third major influence which more or less complements the first two has been our growing acceptance of the 'process' syllabus, though whether this is truly a syllabus in the sense of other syllabi or really a procedure is a question in my mind.) Perhaps in the sense that it is, as Michael Breen says, a plan relating to the major decisions which teachers and learners need to make during classroom language learning which then draws upon the bank of classroom activities and tasks of which I spoke earlier, we may indeed accept it as a syllabus. (Breen, 1987).

It is unquestionably the three foregoing factors i.e. disappointment with outcomes, disillusionment with the succession of the latest 'in' approaches, and the necessity of involving the learners in the decision making process if we are taking their contributions seriously at all, that has predisposed us to make at least a start in applying the process model to our teacher education courses. In this model Breen suggests in his state of the art article on "Contemporary Paradigms in Syllabus Design" there are questions regarding three important aspects of language work which require shared consideration by teachers and learners viz.

(a) questions concerning participation e.g. who works with whom? pairs? small groups? and with whom does the teacher work?

(b) questions regarding procedures e.g. which particular activity will we undertake? how? what resources? for how long? how will we evaluate? and

(c) questions about subject matter e.g. what focus? what learning purpose?

It is these last two which are most often unilaterally decided by the syllabus designer. The Process syllabus, however, provides teachers and learners with "the explicit task of (jointly) prioritising, selecting, subdividing and sequencing what is to be achieved in an ongoing way " (Breen, ibid).

There is little doubt that to those of us used to a more prescriptive and teacher-centred model the process procedure appears threatening and perhaps at first sight rather unstructured. But a sound case is made for it particularly in relation to two very practical situations with which there cannot be a teacher here unacquainted.

Firstly, no classroom group is ever working through one syllabus; in fact the classroom in most cases provides a meeting place of three syllabi - often there is a pre-planned and sometimes external syllabus which the teacher reinterprets for implementation, secondly there are learner syllabi of all shapes, types and sizes; while the third is the syllabus which is worked out day by day and is the inevitable synthesis of the first two (or is it three?) The Process syllabus is designed to facilitate this synthesis through a decision-making process undertaken by teachers and learners together.

Then secondly this process allows us to cope with the ever changing needs of teaching/learning experience in the classroom. The learners' needs, perceptions problems, achievements are continually changing and developing. Says Breen "The process syllabus is a recognition that any syllabus, however carefully planned, is never worked through as the plan itself proposed because teachers and learners are engaged in a complex process which requires the re-interpretation and re-creation of the plan if it is to be made real" (Breen, ibid) and requires it, one may add, almost daily.

One of the possible snags with such an approach I have already illustrated, and this has been emphasised by several researchers. For example Gebhard, in his discussion of collaborative supervision in his article on 'Models of Supervision' (Gebhard, 1984) has pointed out that there is a difficulty in that the ideal and real are sometimes far apart. "Not all teachers are willing to share equally in a symmetrical collaborative decision-making process. A colleague of mine, from a Middle Eastern country, (he says) remarked that if, as a supervisor, he attempted to get teachers to share ideas with him, the teachers would think he was not a very good supervisor" - a direct echo of our experience at W.A.C.A.E. Nor does one have to be dealing with those from very different socio-cultural situations to experience such reactions. Does this mean that the Process model is inapplicable in certain situations? By no means - it merely means that the approach to such joint decision-making must be more gradual and circumspect. To abandon it and retreat to a prescriptive model would surely be to give away two of our most important objectives before we even start viz. the objective of having autonomous, self analytical and self assessing teachers on the one hand and on the other missing entirely the opportunity to show what we mean practically by working with our students to develop sensitivity and response to their learning goals and strategies; the classic "don't do as we do, do as we tell you" situation. For these reasons I feel that while it is not entirely absent, Breen may have somewhat missed the opportunity in his rationale for the process syllabus to stress the development of the autonomous learner; he does stress the development of the group decision making process. Even more surprising to me is that in their in many ways very helpful report on the results of a questionnaire survey of the Practicum in TESOL, Richards and Crookes asked supervisors or instructors responsible for practicum programmes in a wide range of TESOL (or parallel) courses to respond by, amongst other things, ranking 8 objectives for a practicum as follows (these are the rankings arrived at as a result of the survey):

1 To provide practical experience in classroom teaching
2 To apply instruction from theory courses
3 To provide opportunities to observe master teachers

4.5 To give feedback on teaching techniques
4.5 To develop increased awareness of personal teaching style
6 To develop lesson-planning skills
7 To develop ability to select/adapt materials
8 To become familiar with specific methods (e.g., the Silent Way)

It may be significant that, while mentioning objectives like lesson planning, teaching techniques, applying instruction etc other objectives like learning how to analyse, how to be sensitive to the needs of learners are not mentioned, much less that the skill of involving learners in the planning process might be one of the main outcomes/objectives of the practicum. Nor did the report suggest that in the wide variety of skills mentioned by respondents in their 'open' replies was this even mentioned. (Richards and Crookes, 1988).

The thesis of this paper may be stated very simply. There are two objectives we need to build into our teacher education courses, objectives that may have been underplayed by some of us,

1 The development of self analytical, self assessment skills to bring about autonomous lasting growth and development in teachers.

2 The ability to be sensitive to and take into account the needs and objectives of learners and to involve them in a joint planning process in our courses.

If these objectives are accepted as valid then it is the contention of this paper that while there may be other ways of trying to achieve them the most direct and effective way will be by the application of the same objectives and procedures to our own teacher education courses i.e. working not to a prescriptive model - - 'we know how you should teach language if you're going to teach it well' but through a process model which negotiates the content and procedure of the course on the basis of their needs, their perceptions, and their learning styles.

It will be obvious to you all that the procedures I now describe are far from fulfilling that ideal; instead, they are the first tentative steps towards such a model, the first steps towards involving course participants in planning, and in reassuring ourselves that the process is worth pursuing. The situation in which my colleagues and I are working is as follows: We are responsible for a post graduate programme in TESOL; post graduate in the sense that almost all participants are degreed and trained teachers, though they have not necessarily undertaken previous studies in TESOL. The group of participants, having completed a common study on principles of language analysis is divided into three strands corresponding to their intentions regarding the areas in which they wish to teach on completion of the course, in fact almost all are already teaching in these areas or have done so in the past, but without specific qualifications. These areas are teaching English as a second language to adult migrants, teaching English as a second language in schools, and thirdly teaching English as a foreign language i.e. in a non-English speaking environment, though this latter group includes the teaching of English to overseas students who come temporarily to Australia for the purpose, at least in the first instance of learning English.

Adding to the interest of this experiment is the fact that each of these three classes is working not only towards a qualification awarded by the College but is also concurrently working towards an externally awarded Diploma. The strands respectively work to obtain one of three Royal Society of Arts Diplomas i.e. the Diploma in Teaching English as a Second Language to Adults, the Diploma in teaching English across the Curriculum in Multi-lingual schools, or finally the Diploma in teaching English as a Foreign Language to Adults. The reasons for adopting these external diplomas (for those who want them) concurrently with the College qualifications need not be discussed here. Suffice it to say that this places us in the all-too-familiar situation of many teachers where the curriculum is not entirely under their own control but is constrained by an external syllabus, external examinations or the like. It is sometimes argued that this effectively rings the deathknell for a process procedure but we have not found it to be so and would agree with Breen's contention that "the Process syllabus can be appropriate to such a situation because it addresses two of the major problems entailed in the implementation of an external syllabus; how to relate such a syllabus to the internal syllabus of a group of learners and how to gradually create the classroom syllabus of that group which must be a synthesis of external and learners' syllabuses." (Breen, 1988)

We decided that a suitable starting point for trying out these procedures would be the teaching practicum which forms part of each of the courses. There were four reasons for choosing to start with the practicum

1 Whatever arguments there may be about what elements should be found in teacher education courses for TESOL, the practicum is virtually universally agreed upon and identified, particularly by participants, as the most crucial part of the course.

2 Paradoxically the practicum is also the element with which most dissatisfaction is expressed both on a practical and theoretical level. Marion Williams summarised the problems well when she wrote:

> "Classroom observations have, however, always presented problems for teachers and trainers, and generally cause considerable stress and upset on the part of the teacher. Implicit in the approach are various other assumptions: that teaching pedagogy is something that can be both taught and learnt; that observers can tell what is 'good' and 'bad' in a classroom according to some prescribed checklist; and that telling teachers what they are doing, 'right' and 'wrong', will in fact lead to better classroom teaching
>
> Even if one believes that doing this will lead to better teaching, one must ask whether this is in fact the best way of achieving better teaching, and whether individual teachers can and should teach in different ways, in different classroom situations.
>
> (Williams, 1989)

3 The fact that in our particular type of in-service course the Practicum is almost invariably carried out in the teachers' classrooms with students with whom they are very familiar and with the curriculum to a large degree under their own control meant that this was particularly suitable for experimenting with negotiated work.

4 Fourthly, we had the stimulation of much interesting work which had been carried out here in Singapore and reported by Marion Williams under the auspices of the British Council in conjunction with the Ministry of Education, in which a developmental view of classroom observation was posited as against the traditional prescriptive types of supervision which teachers find so threatening and very much at odds with the pupil centred view of teaching which our theory professes. (Williams, ibid)

Our objective therefore was to move from a teacher-educator centred model, not merely to a trainee-centred model, though certainly much more attention would be paid to their views and needs, but ultimately, through them, to a pupil or learner centred model with the focus on the classroom. (I should say that I am concentrating in this paper on the procedures as they affected the teachers working in the 'schools' strand of our courses; my colleagues will later undoubtedly report on the full project, but this will give us a manageable starting point).

STAGE ONE: DISCOVERING PARTICIPANT-PERCEIVED NEEDS (First survey.)

At an early stage of the course, therefore, after about two or three weeks, participants were invited to list those areas of classroom practices and teaching for which they felt the greatest need of help. This timing was chosen to strike a balance between possibly too great a degree of disorientation by complete beginners as against waiting so long that the thinking of participants might, even if unconsciously, have been directed by the course content discussed in curriculum classes. The first sessions of the course had been occupied by linking the curriculum work to be considered with previous studies in language analysis and beginning to look at analysis of needs. Three features marked that first statement of needs by the participants:

1 First, and perhaps not surprisingly in those early days of the course, the range of topics mentioned was very wide. They covered many areas of classroom management e.g. working with groups, getting planned tasks completed. Then there were fairly basic teaching skills applied to the TESOL classroom e.g. questioning, giving instructions, creation of and use of visual materials. And finally there was a long list of specific content areas with which help was wanted e.g. reinforcing new structures, enriching vocabulary, tense continuity, unfamiliar sound patterns, teaching poetry etc. The wide range meant that it was not easy to discover foci for the follow-up work - only a small number of needs were mentioned by several of the participants.

2 The second feature was the interestingly clear emergence of a quite different group of concerns from the responses of participants in other strands of the course. There has been over quite a few years considerable discussion on whether too much distinction may have been made between different areas of TESOL, teaching ESL as against EFL, teaching adults as against children and so on. There have been suggestions of unnecessary distinctions and indeed perhaps even of tendencies towards empire building in making these distinctions

Without any axe to grind and without the desire to exaggerate (there were obviously many skills that were mentioned by all groups) I have to report that quite clearly a different group of concerns emerged from the replies of the TESL in multi lingual schools participants, and notice that these emerged long before any course influence could have affected them. Specific to this group were repeated mentions of language work associated with the mainstream areas of school curriculum e.g. mathematics or social studies activities and management skills arising from this particular situation e.g. how to handle the withdrawal of a group of ESL students from a mainstream class or (more commonly) the skills and techniques required to work as a resource ESL teacher in a mainstream class, team teaching, principles for the grouping of first and second language learners etc. Already this trend, (to become much stronger in later responses and discussions) was evident, even at this early stage.

3 Thirdly, more negatively, except implicitly in some responses about grouping and mainstream needs the emphasis could be said to be largely teacher centred: techniques, strategies, skills, and while it was not entirely absent there is little emphasis on sensitivity to pupil needs and none at all on involving learners in planning and decision making.

STAGE 2: FIRST STEPS TOWARDS MEETING THE NEEDS

After discussion with the course participants the five most frequently mentioned needs were selected for consideration through a modified version of micro-teaching work (Notice the compromise with our process procedures ideals. Certainly the participants had been allowed to participate in decisions - indeed largely to control them - on the subject matter and learning purposes, but they had comparatively little say on the procedures to be followed.) The five areas selected were

1 Giving instructions
2 Catering for the needs of mixed ability groups
3 Meeting the demands of mainstream work
4 Questioning techniques
5 Reinforcing and practising new patterns and structures.

The participants were divided into groups of about seven who followed this procedure:

- The participants chose their own groups and topics.

- Out of each group two participants volunteered to 'teach' their peers.

- The 'teachers' were expected to teach for about 10 minutes with prepared lesson notes. They would explain to the class

  (a) the content; purpose of the lesson, what preceded/was to follow this extract

  (b) the roles they wished their peers to take - crucial in such situations as 'mixed ability groups' and in all cases in indicating levels of proficiency, age, previous knowledge etc. (It was noted that these roles were faithfully and often enthusiastically adopted!)

- The first stage in each session consisted of the other six members of the group (in the absence of the 'teacher' about to give the lesson) discussing and formulating a preliminary list of criteria they would look for, for the skill under discussion.

- The 'teacher' then set and taught the lesson, with the class taking the roles assigned to them.

- The tutor took little or no part in these procedures, certainly not in any directive sense and contented herself with facilitating and videotaping the lesson and interaction/discussion.

- Following the teaching the group spent approximately thirty minutes discussing the lesson in order to establish a list of criteria for the skill (again the tutor took little directive part in these discussions.) Thus, for example, under the heading of 'giving instructions' the group arrived at this list after the first lesson and discussion.

  1 Give precise instructions.
  2 Give clear instructions
  3 Demonstrate visually
  4 Gain pupils' attention
  5 Give pupils a purpose for the activity
  6 Voice - stress the important word in the sentence
  7 Allow for repetition

8 Cater for individual pupils
9 Monitor pupils' completion of task
10 Check pupils understand instructions.

The procedure was then repeated with the second teacher minus the initial discussion, as the criteria established from the first lesson served as the check-list for the second. The criteria established after the discussion on the second lesson (the video tape was used to refresh memories of precise strategies and events) were usually a refinement and extension of the first list into more detailed points e.g. points added to the above list included the helpfulness of rephrasing instructions, using body and gestures to clarify meaning, not speaking too quickly etc.

This second stage was then concluded by the preparation from the criteria established by the group of a set of questions which could be used by participants when planning for a lesson and when evaluating their own performance. The questionnaire could also be used by the visiting supervisor when discussing lessons. Thus it would be the group's own criteria which would be used rather than any externally imposed evaluation of 'good' or 'poor' behaviours.

STAGE 3: REASSESSING NEEDS AND PERCEPTIONS

With stage three we entered the second cycle of negotiation and decision making. Half way through the course the participants were asked again to list those areas of classroom practice and teaching for which they now felt the need of further work and help.

Again three features marked the selection by participants, now quite familiar and comfortable with the procedure:

1 There were many basic similarities with the first list. Many of the skills then mentioned were repeated but with this difference: there tended to be a much narrower focus, a stating of the topic within very specific settings e.g. in repeating the topic 'giving instructions' the proposed setting now was 'Giving instructions to absolute beginners in the language."

2 The first major difference from the initial selection was that the second list was very much shorter and less wideranging. The trends observed in the first selection had become much more pronounced half-way through the course. They were concerned with working across the curriculum i.e. in the mainstream classes whether it was with group work, assessment or in team teaching situations. In other words the differences from the selections of the EFL strand became even more obvious. The two situations and needs are viewed by participants as being quite distinct and with different requirements.

3 The second distinction was the emergence for the first time of perceived needs in analysing, assessing and responding to the needs of bilingual children. The needs of pupils at last emerged as a factor to be genuinely considered though still not to the controlling degree that espousers of process procedures might have wished.

STAGE 4 then saw a repeat of the micro-teaching cycle which proved to require little modification accept in purely technical matters (e.g. better sound recording of the group discussions) from the initial procedures.

Once again the set of self analysing and self assessing criteria were transformed into question form to be used as major areas of concern and emphasis within lessons being taught by participants in their own classrooms and observed by supervisors. The procedure followed was that in planning for a particular lesson the course participant would indicate in addition to normal objectives and procedures the particular focus area which he or she wished to concentrate on in their specific lesson. Participants were encouraged to use the criteria questions in planning for their teaching, and initial pre-teaching discussions with the course tutor/supervisor focussed on how the criteria had been applied in the plan.

Notice that to this point of the experiment participants have for obvious reasons, been encouraged to focus on only one of the selected criteria areas (e.g. catering for the needs of mixed ability groups) for any one lesson. Post lesson discussion with the observer while not concerned exclusively with the selected area have certainly had this as the major area of concern each time, with the candidates being encouraged to analyse and assess their own performance in the light of their own criteria.

SUMMARY OF CONCLUSIONS

What have we learned from this still incomplete attempt to apply process procedures to this element of a teacher preparation course in order to develop self analytical and self assessment skills in the participants?

While findings are largely impressionistic at this stage and must therefore be expressed tentatively and cautiously, the following seem to be emerging:

1 The process procedures involving participants in decision making and planning for these elements of the course work appear to be bringing about far more effective changes in classroom teaching behaviours than our old prescriptive or transmission model ever did. While we would like to believe that this is because of the procedures adopted we have to be cautious about this conclusion as the possibility remains that this could arise simply from the more intensive work carried out on specific elements of classroom procedures.

2 In spite of the novelty of being consulted on subject matter and the construction of their own criteria, course participants undertook these procedures with apparent ease and competence, arising no doubt from the fact that they were all experienced teachers accustomed to the decision making process, though perhaps not in this particular context.

3 Nevertheless the repeated warnings by previous investigators that the process procedure was not equally welcomed by all participants proved true; a small minority still prefer the prescriptive rather than investigative model.

4 One unquestionable outcome has been the lowering of the discomfort levels so often reported by participants in association with classroom observations by supervisors. With the transformation of this element from an assessing prescriptive approach into an investigative, collaborative and self analytical procedure using their own criteria, participants consistently report that what was previously 'teaching supervision' has become much less threatening and conversely more useful as a developmental and cooperative process.

5 One small almost 'side' discovery was the usefulness of role-play in the modified micro-teaching procedures. While the weaknesses of peer rather than pupil teaching in micro-teaching have often been discussed, in this context the requirement for participants to imagine themselves into the particular roles assigned to them by the 'teachers' in their groups was several times referred to

later in teaching sessions as helping with building sensitivity to needs and procedures that relate to particular students. A couple made it clear that they had particular students of their own in mind when carrying out the role-play.

6 If there is an element of weakness in the procedures described it lies in the gap between the participants' perceptions of their needs and problems and the perceptions and needs of the language learners themselves, their pupils. There is an as yet not entirely bridged gap between what some investigators have described as a problem solving approach and one based upon classroom decision making and investigation (Breen, Candlin, Dam and Gabrielson, 1989). Certainly the link between the micro-teaching sessions and actual tutor/participant discussion of samples of classroom teaching is helping to bridge that gap but it would be good to see the classroom situation and learner needs become the focus of course work submitted by participants to a greater degree.

7 The question of whether the process model is helping to produce more autonomous teachers who are themselves willing to involve their learners in the decision making process: these questions remain at this early stage open, though we might without being accused of exaggeration say that for at least the first part of the question the signs are good.

8 Finally the positive response and apparently favourable outcomes of the adoption of a process model for this element of our courses may point to the possibility that a larger proportion or perhaps even the whole course could profitably be constructed on this Model.

[I should like to express my deep indebtedness to

- my co-researcher, Elizabeth Valkhoff, who helped plan and design this research and who will undoubtedly report later in this or another forum on the full project and in particular on her findings with the EFL group.

- our research assistant, Rowena Errington who conducted and analysed the micro teaching sessions.

- Michael Breen, whose presence in Western Australia this year undoubtedly proved a stimulus to the developments reported here, as evidenced by the many references in this paper.

- to the Director and Board of the Institute of Applied Language Studies, Western Australian College of Advanced Education, for providing the funding necessary to try out and investigate these procedures.]

REFERENCES

Allwright, R.L. (1972) "Prescription and description in the training of language teachers" in Qvistgaard et al (eds) Applied Linguistics: Problems and Solutions, AILA Proceedings, Copenhagen 3: 150-166.

Allwright, Dick (1988) Observation in the Language Classroom, (Longman)

Breen, M.P. (1987) "Contemporary paradigms in syllabus design, Parts I & II, Language Abstracts, April and July, C.U.P.

Breen M., Candlin C., Dam L. and Gabrielsen G., (1989) "The evolution of a teacher training programme" in Johnson, R.K. (ed) The Second Language Curriculum, 111-135, C.U.P.

Chomsky, N. (1966) "Linguistic theory" in Mead R.G. (ed) Language Teaching: Broader Contexts, Northeast Conference Reports.

Felix, S.W. (1981) "The effects of formal instruction on second language acquisition" Language Learning 31(1): 87-112.

Gebhard, J.G. (1984) "Models of supervision: choices" TESOL QUARTERLY, 18(3), 501-514.

Long, M.H. (1983) "Does second languagc instruction make a difference?" TESOL Quarterly 17(3): 359-382.

Politzer, R.L. (1970) "Some reflections on 'good' and 'bad' language teaching behaviours" Language Learning 20(1): 31-43

Richards J.C. and Crookes G. (1988) "The practicum in TESOL", TESOL Quarterly, 22(1), 9-27.

Sampson, G.P. (1984) The Educational versus the Scientific Bases of the Communicative Approach. Confcrence Paper, SEAMEO

Williams, M. (1989) "A developmental view of classroom observations" ELT Journal 43:2, OUP.

BRINGING EVALUATION AND METHODOLOGY CLOSER TOGETHER

Davia Crabbe

1 INTRODUCTION

In this paper I address the problem of learner evaluation of methodology and, in particular, task-based methodology. At the end of any course it is almost required practice to give a questionnaire that elicits learner evaluation of aspects of the course. This is often the only formal evaluation of the course that is carried out. I have distributed many end-of-course questionnaires, but on more than one occasion I have felt that they somehow miss the point. Firstly, there is, typically, a paucity of information that comes back from them. They only give a sense of the degree of client satisfaction or dissatisfaction which is commercially but not pedagogically informative. It may well be that the design of these questionnaires is lacking but if client satisfaction is the main thing that comes out of them, the end of the course is a little late to gauge it. Secondly the comments made are usually one-off statements that can easily be dismissed as the eccentric whim of one respondent. Sometimes one would like to be able to discuss a viewpoint with a respondent because one feels that there is a serious lack of shared expectations of the course, something that should have been ironed out in a different way. Clearly a final questionnaire does not do justice to the depth of involvement in the pedagogical decisions one would like students to demonstrate.

This paper begins with the assumption that course evaluation by learners is based on self-assessment of their own communicative performance. A general sense of progress is likely to be attributed, at least in part, to course activities. What is suggested here is that evaluation based on a general sense of progress is not good enough if self-direction is a goal. Evaluation of methodology needs to be encouraged right from the beginning and to be focussed on specific tasks. If learners see value in tasks they are more likely to use them for independent work. The paper proposes certain requirements of task design that may help this important process. Any final questionnaire should reflect how far this process has worked.

As with much teaching procedure, the proposals are somewhat speculative in character. They arise out of experience with a task-based pre-sessional course for overseas postgraduate students run at Victoria University, Wellington. One of the four general aims of this course is for the learners "to know what steps they might personally take for further improvement of their communication in English in an academic context." It should be added that the students on this course are highly motivated by their imminent need to survive in demanding academic contexts.

2 THE LEARNER AND EVALUATION OF TASKS

When people are learning a new skill, my experience is that they are informally evaluating a great deal - evaluating not primarily the effectiveness of the task for learning but their own performance in doing it. If you are learning to ski you are constantly critical of what you are doing, that you are leaning at the right angle to the slope, that your feet are appropriately placed and so on. This is a natural informal evaluative process that one would expect of any skill learning, including language learning.

What I would like to suggest is that other evaluation arises largely out of this self-assessment. If you are being instructed at skiing and you are not told about angle to the slope until you have made several undignified slides downhill in a prone position, you will probably evaluate negatively the way in which your practice task has been explained to you. In the same way, the learner of a language is likely to judge classroom tasks against the progress that he perceives he is making. If he perceives improvement or no improvement, he will either attribute that improvement or the lack of it to the classroom tasks, or to his personal study and use of the language, or to both.

It seems to me that it is important that the learner is able to distinguish what contribution the classroom tasks on the one hand and personal study and general use of the language on the other, have made to his language learning. The reason for this is that the classroom tasks are in the public domain and the personal study is in the private domain and in the interests of establishing self-direction, the boundaries between these two domains, usually raised by education, need to be removed. In other words, the language learner needs to be evaluating across both domains which aspects work and which do not so that he has control over them. In this way, classroom tasks become a source of information, a model, for personal study (and vice versa) and even a model for managing general use. Classroom tasks should therefore highlight and not just require efficient strategies for language learning and use.

This seems an obvious criterion for task design when self-direction is an objective. Yet it is not often a criterion which is met by classroom tasks. If we are to help a learner to evaluate a classroom task for the degree to which it enhances performance and learning and, in so doing, to help him to build up a personal arsenal of independent activities, then I think that certain requirements have to be met by classroom tasks in a course. I have listed these below and each will be discussed in turn in subsequent sections. Considerably more attention will be given to the third requirement.

(1) Tasks should be identifiable by learners as involving a specific piece of communicative performance

(2) Tasks need to be done in such a way that they can be easily staged by a learner working on his own.

(3) Tasks need to include an element of enhanced feedback and practice to demonstrate improved performance and thus facilitate evaluation.

3 TASKS SHOULD BE IDENTIFIABLE BY LEARNERS AS INVOLVING A SPECIFIC PIECE OF PERFORMANCE

If learners are to be able to attribute improvement in language development to any particular activity, a general sense of improvement is not easily attributable in a valid way. I know if I practise at the piano simply by playing it as often as possible, and I improve, then it is difficult to say why, beyond the fact that I have played a lot. If, on the other hand, someone says to me that I need to focus on my fingering and shows me a technique to practise that aspect of playing the piano, I can say the technique was either useful or not depending on whether there is an immediate improvement in fingering. In the same way, if I am using and studying the language a lot and gradually improve, I do not know what to attribute that improvement to. It may be vocabulary study, it may be extensive listening to the radio, it may be both, it may be neither. Because I have not paid attention to any specific task and focused on that, I cannot really tell. Does it really matter, so long as we get there? Well I think it does matter. For one thing, it may be more efficient to concentrate on one bit of performance at a time on the grounds that an in-depth case study of a bit of language in use is better than trying to draw generalisations from language data spread over several bits of performance. This view draws on an information processing view of language (McLaughlin 1987, Chap. 6) rather than the comprehensible input view of Krashen. For another thing, it gives the learner a greater satisfaction with the learning process in that he can evaluate specific progress as it happens on one front rather than have a vague sense of general progress on several fronts at the same time.

This is really an argument for task-based learning in general - not any tasks but tasks that have as their goal a specific bit of performance with high face validity, that is performance that relates to the learner's target communication. This bit of performance might be writing a particular genre of report, being interviewed for a job, giving a seminar presentation. The performance may of course be considerably guided or simplified for the level of the learner using various techniques available. (See, for example, Phillips 1983 for the principle of reality control and Widdowson 1979 for the technique of gradual approximation).

This suggestion, that specific performance tasks may enhance self-assessment and evaluation of task effectiveness, implies a problem with tasks such as reordering jumbled sentences, some information transfer tasks, spotting the difference in two pictures, and many information gap activities. These kinds of tasks are stock in trade for communicative language teaching and whilst I do not wish to decry their value for fluency development, they are often tasks that are not specific in performance. What they usually aim at is general improvement in proficiency and this provides no performance focus for learners to evaluate except artificial classroom performance. While learners can evaluate their progress in these artificial tasks, such progress may be seen as trivial. More importantly, however, they provide no encouragement for the learner to relate the classroom task to the real tasks that he will face or is already facing in the real world. He is therefore, I believe, less likely to identify or consciously transfer learning strategies.

The current ESP practice, then, of simulating target communication in the classroom through tasks such as essay writing, preparing and delivering seminar presentations, listening to lectures, is valuable not only because they meet Phillips' criteria of non-triviality and authenticity (Phillips 1983) but also because at the same time they meet one requirement for training learners to meet real learning needs themselves. If a learner in the target situation has a problem with oral presentations, then will his mind turn to an information gap activity to improve his oral performance? Probably not, but if he had done a task involving individual preparation procedures for a short talk he would then have a model procedure to follow.

4 TASKS NEED TO BE DONE IN SUCH A WAY THAT THEY CAN EASILY BE STAGED BY A LEARNER WORKING ON HIS OWN.

This requirement is a practical one. I think that not infrequently on English language courses, including EAP courses, tasks are selected that involve special materials or special classroom equipment or special management (information distributed in a certain way, for example). Again, I do not wish to decry the innovative design that is behind many of these activities, but I do believe that it may mystify the teaching process by making the teacher the powerful magician. This may prevent the learners from evaluating a task as something which they can usefully stage themselves. The boundaries between the public and private domain remain intact. Students - and teachers - often believe that purpose-built materials are necessary for language learning. I am not suggesting a contrary minimalist approach, but purpose-built strategies are so much more important, and so much more difficult to provide.

On the EAP course in Wellington, naturalistic performance is emphasised as much as possible in the sense that tasks are mostly tasks that can be done either in groups or individually without any extra resources. The fashion for group work, supported as it is by work demonstrating the quality of the interaction involved in such work (Long and Porter 1985) tends to overshadow the arguments in favour of individual work in EAP where the conceptual performance is intimately bound up with the communicative performance and in the end the learner is on his own. (Crabbe 1987)

5 TASKS NEED TO INCLUDE AN ELEMENT OF ENHANCED FEEDBACK AND PRACTICE TO DEMONSTRATE IMPROVED PERFORMANCE AND THUS FACILITATE EVALUATION.

One of the biggest problems that I see with many communicative tasks, even if they concentrate on specific performance, is that they obviously provide for communication but they do not obviously provide for learning. Of course the current theory is that communication does lead to learning and the consequent principle is that the more communication you do, the more learning that will take place. This principle is not always accepted by learners. They worry about their performance - about not understanding bits of the communication, about making production mistakes. Moreover their fears about their performance means they feel they are not making progress. This is likely to have negative repercussions. The students are likely to undervalue the course as a whole and, moreover, no particular task will stand out as one they can take into the private domain as independent language learning strategy. I believe that there is an element of task design that is critical here and to illustrate this I want to describe two language learning demonstrations I experienced, one 20 years ago as a beginning student of Russian and the other 10 years ago at a seminar in Lancaster.

The Russian course I attended was a traditional grammar and translation course with a bit of audio-lingual laboratory thrown in. Once a week we had a Russian evening in a local cafe and on one night a member of the Russian diplomatic staff came along to engage in conversation. He chose to play a recording of Goldilocks and the Three Bears in Russian perhaps to avoid holding what must have been painful conversation with us. Then he had us retell the story, person by person, one sentence from each person. If the person did not get it right the turn passed to the next person. The recording was replayed between each retelling. I thought this was a marvellous way of learning as we each struggled with our sentence, not only with the form but also with the content. We had to do it several times and every time round we each got a different sentence to do.

The seminar at Lancaster was held by Celia Roberts, at that time teaching English to Hong Kong policemen. She demonstrated a piece of performance which was answering a telephone at the police station. She played the part of an irate member of the public phoning in to complain about the noise in the port area. She held an imaginary phone to her ear, made a ringing noise and pointed at a hapless member of the audience to answer. The answer was inappropriate and so she hung up and started again pointing at different participants and hanging up until a correct or appropriate response was made. Each participant learned from others mistakes although there was no explicit feedback.

What are the features of this and the Russian task? Firstly the feedback is built into the task. Getting the performance right is an essential criterion for it to be completed. Sometimes there was an explicit model for comparison as in the Goldilocks example, sometimes the feedback is by listening to other speakers perform or getting correction, implicit or explicit, from the teacher.

The second feature is that there is what I call repeated performance. In other words, learners get a chance to have another crack at the same communication, not another piece of communication with some similar features. I think as we emerged from the audio-lingual era we forgot the importance of repetition in language learning, so keen were we to slough off the old paradigm. Of course unlike the audio-lingual repetition of unconnected forms, I am talking here about the repetition of connected meanings, discourse.

In the tasks on the Wellington EAP programme, there is an attempt to incorporate as much as possible the features of built-in feedback and repeated performance. An example of a ready-made task that can be used with any content available is the 4-3-2 technique (Maurice 1983). In this task a learner gives a talk for 4 minutes to a partner and then listens to the partner give a talk on the same topic, a new partner is found and the same talk is given and listened to but this time for 3 minutes. Finally a third partner is found and the talks are given again in 2 minutes. Feedback comes from listening to others give the same performance but to enhance the feedback process I would deliver a talk myself on the same topic as a native speaker model after the 4 minute and the 3 minute talks, thus providing a native-speaker model for performance comparison. The topics are usually drawn from the study theme that was currently being worked on. The activity can be used at any point for practice in oral presentation.

Writing tasks also involve repeated performance and built-in feedback as the writing is done in the form of a workshop where the learners work independently calling on a teacher as informant when needed. On the same principle, when giving a short seminar style presentation the students are encouraged to practice the presentation several times on their own at home before they deliver it. Feedback is provided after the presentations and the learners are able to benefit from the feedback given to others before their own turn comes around.

A similar and more fully developed approach to task design, is described in Willis and Willis (1987) in which there is emphasis on rehearsal of tasks and on listening to native speaker models. There is less emphasis on the repetition of the same performance although this is not precluded.

I think that there are important reasons why built-in feedback and repeated performance are necessary components of any task and this is to do with the nature by which we learn. Built -in feedback enables the learner to critically assess how well he has performed, not in general but in specific details. The best way in which this is to be encouraged is still not clear although I favour procedures by which the learner has to discover the errors for himself (see Chaudron 1987, however, for a review of error correction by the teacher). Repeated performance enables the learner to apply the results of feedback as well as develop a degree of automaticity. After a number of repetitions, students nearly always report improvement, at least in fluency. This gives repeated performance high face validity with learners - there is hard work involved in repeating a bit of communication four or five times but the perceived return avoids any sense of tedium. Some research to support the perception of improved performance is that carried out by Brown and colleagues (Brown et al 1984) in which little performance improvement in oral tasks was noted after simple repetition but when the speakers had a chance to listen to others do the same task, there was significant improvement in subsequent performance. This research was with native speakers. Arevart (1988) looked specifically at the 4-3-2 technique with second language learners and found that their fluency increased and that "repetition also results in improvement in the accuracy of the language used in the talk. The case studies show that the learners correct grammatical errors previously committed while speaking. They set out a discourse plan formulate utterances, establish language rules and try them out." (p 80) This happened without a native speaker model for comparison.

Clearly research is needed to confirm that any gain in performance in such tasks is permanent. At this stage however, I am satisfied with the fact that immediate improvement is evident to the learners themselves. The learners are actively engaged in managing improvements and I believe this helps to break down the public/private boundary. The task is more likely to be transferred to the private study domain as a useful strategy for practice.

In the private domain there is of course the problem with feedback in self-directed productive tasks. A model will usually provide feedback of a comparative nature for learners to identify salient lexical, structural or even pragmatic information for their own personal learning. However models are not available for tasks that you do without a teacher unless you are working with specially prepared materials with models built in (See Willis and Willis 1987). But tasks can be made out of models. In other words a learner can take a piece of available native speaker performance (printed or recorded), put it aside as a model, do the performance himself and then pick up the model again. Classroom tasks may have to reflect this order of going about things.

6 THE WIDER CONTEXT

Designing tasks that assist evaluation of effectiveness, do not of course represent the whole picture of how students are encouraged to evaluate - either their own performance or the effectiveness of the programme. Evaluation, to be effective, involves the gathering of information, the coding of it in a way that is sensible and usable and then applied. All this is hard enough for a teacher to do of his own performance. It is extremely difficult to encourage the learner to do on their own behalf. Yet unless we take this on as part of our job and particularly so in the case of EAP courses where the learners will soon be fending for themselves without formal language instruction, there is little hope for any evaluation of the course being very meaningful except as a measure of client satisfaction. What we want those final questionnaires to reveal is not that the learners liked the course, although that is important enough, but that they were so involved in the learning experience that they knew what was going on.

The Wellington EAP course referred to here used a wide range of strategies, embodied in minimum standards for self-direction, to encourage learner evaluation. These strategies included student record booklets distributed at the beginning of the course and completed by the students as the course progressed, personal interviews, explicit discussion of tasks, an introductory study theme on how people learn languages, a self-access centre with advisors and, in all of that, an attempt to foster metacognitive awareness of learning. Even with all that effort one cannot be sure that self-direction is developing. The affective aspect is another factor in the process, perhaps the biggest factor of al! and although that can be addressed through continuous monitoring of individuals, there are always limitations.

7 SUMMARY OF MAIN POINTS

In this paper, I have made the following claims

7.1 That final course evaluation questionnaires do not reveal a great deal of information about the course.

7.2 Part of the reason for this is that the learners do not usually have much basis for attributing improvement to any particular aspect of the course.

7.3 That learners should therefore be given such a basis as it helps them to be self-directed and to transfer tasks from the classwork to personal study

7.4 That, to achieve this basis for evaluation, three requirements of task design are that

(i) tasks should involve one specific piece of performance so that improvement in that performance is attributable to that task

(ii) class tasks need to be stageable by learners on their own. All group tasks should also be performable as individual tasks.

(iii) tasks need to include repeated performance in order to enable learners to evaluate progress in specific performance and thus to increase their management of learning.

7.5 That these requirements of task design constitute one aspect of a broader strategy to develop self-direction.

8 CONCLUSIONS

I have said that if we evaluate tasks for whether learners can evaluate them positively for effectiveness, I think it would be surprising how often the tasks do not measure up to this criterion. In the end, of course, we should as teachers know more than the learners about learning and a traditional view would be that the learner should trust us. I think it would not be too difficult however to make our wisdom about learning, such as it is, more transparent and accountable so that learners can take it on and not need us when we are not available.

This suggests that in addition to research questions designed to evaluate which aspects of tasks seem to contribute to performance improvement, we need parallel research questions to evaluate how visible this performance improvement appears to learners and whether high visibility leads to transfer of strategies.

REFERENCES

Arevart, Supot 1988. The effects of repetition on spoken fluency and accuracy. Unpublished M.A. thesis, Victoria University of Wellington.

Brown, G., A. Anderson, R. Shillcock and G. Yule 1984. Teaching talk: strategies for production and assessment. Cambridge University Press.

Chaudron, Craig 1986 The role of error correction in second language teaching. In B.K. Das (ed) Patterns of classroom interaction in Southeast Asia. pp 17 - 50 SEAMEO Regional Language Centre (Anthology Series 17)

Crabbe, David 1987. Developing conceptual performance in a foreign language. Paper presented at RELC Regional Seminar.

Long, Michael and Patricia Porter 1985. Group work, interlanguage talk and second language acquisition. TESOL Quarterly 19,2.

MacLaughlin, B. 1987. Theories of second-language learning. Edward Arnold.

Maurice, K. 1983. The fluency workshop. TESOL Newsletter, 17,4

Phillips, Martin K. 1983 Towards a theory of LSP methodology. In J.Wilson and P. Horey (eds) ELC Occasional Papers No 1. King Adbulaziz University.

Widdowson, H.G. 1979. Gradual approximation. In Explorations in Applied Linguistics. Oxford University Press.

Willis, David and Jane Willis 1987. Varied activities for variable language. ELT Journal 41,1

EVALUATING A TEACHER TRAINING PROJECT IN DIFFICULT CIRCUMSTANCES

C J Weir and J Roberts

This paper presents a description of the procedures adopted in a recent evaluation of the effect of a teacher training programme on student language performance. It is hoped that the account will lead to constructive discussion of how to improve the methodology employed. It considers the problems that may be faced by external evaluators working in difficult circumstances. We are grateful to the Overseas Development Administration in the United Kingdom for giving us permission to report our methodology.

1 BACKGROUND

The SEPELT Inset Project in Nepal was set up to provide 1080 standard 8-10 (Upper Secondary Level) English teachers with one month's inservice training, delivered by locally trained Nepali staff, working from a standard course manual and supported by an expatriate training officer. It ran from 1987 to 1989.

The long term goal of the training was to improve students' performance in the School Leaving Certificate English examination. The course provided training in basic ELT procedures designed to enhance the teaching of the National English Curriculum.

The Nepal baseline study described in this paper was a small scale, field based, non equivalent group study, contrasting the learning gains of students in the grade 8 classes of 11 trained teachers and 11 untrained teachers. The study established procedures for measuring the effect of the SEPELT training on students' language performance. It was also concerned with determining the suitability of these procedures for evaluating similar projects elsewhere.

A small scale non-equivalent control group pretest- posttest design was employed. In this design two groups of students which are similar, but which are not formed by random assignment, are measured both before and after one of the groups undergoes the experimental treatment.

In this case the experimental treatment took the form of instruction by teachers who had attended the SEPELT training course. We were concerned to see if, with faithful implementation of the training, there would be superior learning gains by this group as evidenced by improvement in student language test scores.

As well as testing students we had to monitor the performance of trained (Experimental) and untrained (Control) teachers to establish that the treatments received by the pupils were indeed different, i.e., were our control and experimental groups exposed to different language instruction?

We made short visits to Nepal in November 1988, January 1989 and November 1989. As a result of the first visit the baseline framework was set up and technical staff contracted for data collection (The New Era Research organisation). A short training course was provided for technical staff during the second visit. The final visit was made in order to monitor data collection.

2 The Language Test Instruments

To determine the effects of the training programme, base line tests were administered to the new intakes in grade 8 at a selected group of schools (12 Control and 12 Experimental in the first instance). This took place at the start of the school year in February 1989. These classes were of both the experimental type, where the teachers had been on a training course and of the control type, where the teachers had received no prior EFL training. The tests were readministered at the end of grade 8 in November 1989.

Part I of the battery was constructed to sample as widely as possible the structural elements in the English syllabuses for years 7 and 8 in the Nepali Upper Secondary School System on the basis that such linguistic elements would be accessible to both control and experimental groups. The two parts of this general proficiency section of the battery were prepared in advance of the November 1988 visit and piloted during that visit.

They consisted of:

PART IA Selective Deletion Gap Filling 100 items 1 hour

Passsages were selected and reconstructed from the grade 7 and 8 textbooks used in Nepal. They were rewritten taking care to employ only those structures and lexical items occurring in these course books. Items were then deleted from these passages to sample as far as possible the range of structural items in the national curriculum for these grades. The task for students was to repair the deleted items by writing the missing word on an Answer Sheet provided. A similar technique was employed in Davies et al's 1984 Survey of English Language Teaching in Nepal. The latter differed in that it was a test which had originally been designed for use in Malaysia.

The properties of many realisations of this test method are high correlations with other general proficiency measures and with tests of reading comprehension in particular. We wanted to have a large number of items with a wide range of difficulty because the test might have to show improvement over a period from 1 to 3 years. The problem with the more familiar passage plus comprehension question format is that the number of items that can be set is restricted. By using a gap filling test it is possible to create a far greater number of items which we thought would demonstrate development in linguistic competence.

It was hoped that students would be able to complete some of the items learnt in Grade 7 in the first administration and by the end of the year it was hoped they would be able to score on those set on the year 8 syllabus.

The experimental group might well be expected to outperform the control group on this section of the test as the training course was aimed at improving the teaching of the existing materials in the Grade 8 reader and in the long term raising the number of successful passes in the SLC in Grade 10. The ability of students to produce structural items was an anticipated result of the trainees' implementation of their training. It would still however be a fair test for the control group as it did not contain any structures or lexis that were not present in books 7 and 8. Gap filling exercises are also present in the course books.

PART IB Dictation 30 minutes.

This was a forty item test, again of general proficiency, based on sentences (mainly imperatives, instructions and directions) taken from books 7 and 8.

It differs from test 1A in that as well as correlating well with other measures of general proficiency it has a good record of correlating highly with other measures of listening ability.

It was felt that if this test was to be readministered in 1989 and 1990 then it obviously must not be too easy to start with or otherwise a ceiling effect might negate the possibility of measuring achievement and identifying any increase in scores in either of the two groups over the period of the study.

As with the gap filling test the purpose was to determine whether the experimental group performance would outdistance the control group particularly as the trained teachers were trained to use more English as against Nepali in the classroom and a number of the activities in the training manual encouraged this. So once again this measure of general proficiency was designed to reflect the main purpose of the training course namely to enable the teachers to improve the effectiveness of their teaching.

As tests of general language proficiency, both dictation and gap filling can still be considered fair to the control group as they do not contain any language extraneous to the course materials and to a lesser or greater extent dictation and gap filling will occur in the lessons of both groups. Dictation is advocated in the national curriculum which states that at the end of Lower Secondary Level the student will be able to "take a dictation from any of the prescribed materials in the text book".

To summarise, the expectation was that with improved teaching methods and a greater use of English in the classroom the experimental group would improve at a greater rate and that eventually performance in the SLC would reflect this.

PART II

Another purpose of the discussions with the trainers during the first visit was to identify criterial, behavioural differences that might be expected to emerge in students' performances as a result of this short in service training programme.

We have already commented on the general aim of the training to make teachers more efficient at what they did already and our feeling was that Part I of the test adequately catered for this aspect. Our concern in Part II was to reflect any differences in kind, in terms of student performances that might be expected to emerge from the training.

The training team's view was that it was in the skill of writing that clear differences and new behaviours were likely to occur. The experimental students were more likely to be able to create meaningful new sentences and to execute controlled writing tasks. The control group were more likely to copy from the board and memorise and reproduce paradigms provided by the teacher. They also felt that there might be increased oral interaction among students as a result of the training. However, the considerable difficulties and vast expense of conducting spoken language exams precluded their use under the conditions obtaining in Nepal.

We restricted ourselves to trying to establish whether any differences in written production occured through employing test tasks to relect these activities in Part II of the battery.

During our first visit in November 1988 we had asked trainers and trainees to write a short essay on a topic that would be familiar and accessible to their students in year 8 and also to prepare a framework for a cued writing task on the subject. We thus had thirty five examples to provide us with an idea of levels, topic areas and content. This enabled us to produce four cued writing tasks from which we selected the final two used in the battery after trialling in Katmandhu in December 1988.

3 The Process Instruments

Development of Instruments

Prior to the November 1988 Nepal visit, an inventory of training characteristics was produced from the manual and other available documentation.

In discussion with Nepali trainers during the November 1988 visit, the features that they considered both to be of highest priority and the best discriminators of trained and untrained teachers were selected from the inventory and the resulting short list was then used as the basis for the observation instrument. Additionally this list helped us identify more clearly those training characteristics potentially capable of effects on measurable pupil performance which would need to be reflected in the tests.

A revised teacher self report instrument was produced after discussion with Nepali trainers, and a copy left with a request for piloting.

On return, we trialled a draft observation schedule in a local secondary school's classes of French. The design of the resulting schedule took into account its implementation by technical staff, with an emphasis on simplicity of use and a focus on low inference observations. The revised form was the basis for observer training conducted in January 1989 by Weir (see Appendix 1 for details of the schedule).

Instruments: Observation

On page one of the schedule details of the school, observer, teacher, class and lesson are recorded. On page two the observer must code three five minute samples of classroom talk into categories of: teacher and pupil talk; use of English and Nepali; use of questions as contrasted with all other forms. The samples are selected according to strict time criteria. These data give an indication of the proportion of English used in class, who uses it and to what extent questions are used by either teachers or pupils. On page three, the observer is provided with a checklist of classroom activities, to indicate if an activity has occurred, irrespective of duration. For each activity identified, a short illustrative note is required. These data should indicate the ocurrence of activity types which discriminate trained and untrained teachers. Also, the observer writes unstructured notes to describe the whole lesson in terms of teacher and student activities, including those in progress during coding. While these notes are unstructured, they are based on a previously identified lexicon of appropriate action verbs. These notes provide a "thumbnail sketch" of the observed lessons which can also be cross-referred with checklist and coding data.

On page four, the observer is required to estimate overall talking time in the lesson and the respective proportions of English and Nepali used by teachers and pupils.

It may seem that the data obtained present a somewhat narrow and simplistic account of classroom processes. It should be noted that there is nothing to be gained by making observers' tasks more complex than necessary; that key training and discriminating characteristics can be identified; and that talk in elementary ELT classrooms is intrinsically controlled and restricted in range.

Additional Data Collection

Two other forms of convergent data were required: teachers' self report lesson descriptions and pupils' work.

Self Report

On each visit by technical staff, the teachers were given self report forms. They were asked to describe three recent, typical lessons they had given.

Self report is often considered unreliable, reflecting impression management rather than actual practice. However,teachers are unlikely to report doing what they never do, or what is unknown to them. These data were used, with caution, as additional information on the customary practice of teachers in the two groups.

Pupils' Work

As part of the final observation visit, technical staff were asked to obtain samples of work from about 5 pupils in each class. These data were obtained to help identify discrepancies with observational and self report data.

Teacher Interviews

Structured interviews were held with the 18 teachers who attended a meeting during the third visit. Data were obtained on the following features: years of service in the school; educational background and training; the teacher's place of origin; other occupations; number of pupils in the school; number of pupils in class ; number of lessons per week; number of lessons in the year; school's SLC pass results for last year, both general and in English only; likelihood of teacher continuing with the class in grade 9; an estimate of the teachers' level of oral English, using the British Council ELTS scale.

4 Selection of sample

Location

The design of the study was heavily influenced by the serious problems of communication and information gathering in Nepal. Most schools do not have telephones and postal delivery is highly unreliable. Telegrams can take up to 6 weeks to arrive and the only means of ensuring messages getting through is by personal delivery. These problems were compounded in November 1989 by India's obstruction of key Nepali imports, notably petrol, which made all travel extremely difficult. District Education Office files often do not contain complete or up to date information, such as lists of school staff. Without actually visiting the schools it is not possible to ascertain whether particular teachers are still teaching there or not. As a result a large scale study of a widely dispersed sample of teachers was always out of the question. These factors also indicated the need to employ experienced field workers through the local New Era research organisation rather than ELT subject specialists with no fieldwork experience.

Kathmandhu valley would have been by far the most convenient place to conduct the study, but it is evident that it is quite unlike any other region of Nepal, and would not have provided representative sample schools, particularly since most of the training took place outside Katmandhu valley.

Pokhara region was chosen after discussions with the project leader during the summer of 1988. It was considered to be a fair representation of rural regions outside Kathmandhu, where the main training effort has been going on. 97% of Nepal is rural and 6 districts are contained in the Pokhara region and it includes many quite remote schools. It has relatively good road communications and some of the best contacts with schools and regional directorate of education were in that region. Access to sample schools was possible within a day, (if public transport could get fuel), so greatly reducing the cost of employing technical staff and limiting the overall time spread of test administration. (Given the limited time span of the study the longer it took to administer the baseline tests the less comparable would be the results of the study).

Selection of Teachers

The project leader in consultation with other training staff was asked during the first visit to select 16 teachers who had been trained in either the past or current Pokhara Inset courses.

The main selection criteria we required were:

i) The teachers were thought to be likely to implement their training, in its key characteristics.

ii) The teachers did not work in a school known to be exceptionally different from other sample schools.

The trainers were also asked to select 16 untrained teachers, with the aim of providing a roughly comparable control group. Initial selection was done according to best available local knowledge of the trainers and the local Regional Education Office.

Further selection criteria for both groups were:

i) Pupils and teachers in control (C) and experimental (E) groups should be as equivalent as possible in terms of language ability. Methods to ensure this as follows:

a) Pupil equivalence:
SLC results of (C) and (E) schools should be compared and schools with equivalent scores included in the study.

b) Teacher equivalence:
The Part 1 language test designed for the baseline study was administered to the teachers during the SEPELT trainers' initial visit in January. Teachers with widely disparate language levels were at this stage dropped from the study. This resulted in reducing the n in each group from 16 to 12.

This obviously meant that we were reducing the potential effect of improved English arising out of the training because our main concern in this study was to see if improved teaching methods made any difference to pupil language scores.

ii) There should be no special features in the school intakes which would bias the sampling, eg, extreme variations in parental income, school to school or rural versus urban.

iii) Access to the schools by technical staff should be both possible and welcome.

iv) The teachers should remain with their grade 8 class throughout grade 8 and should be likely to continue with the same pupils through grade 9.

v) There should be equivalent stability in pupil population in both (C) and (E) group schools; that is attrition rates should not differ markedly.

vi) All schools in the study should be well enough organised and run to ensure the efficient collection of test and observational data.

vii) There should be adequate facilities for testing, to minimise student copying

viii) As far as possible (C) group teachers should not receive informal "secondary training" during the period of the study, eg, by contact with trained teachers.

ix) It was essential that (C) teachers should not attend a training until late 1990.

x) Reaching each school should be possible in the period of the study, and very remote schools were to be excluded.

As far as possible an initial selection of 16 untrained and 16 trained teachers was made on the basis of these criteria with the expectation of some scaling down in sample size because much of the information necessary was simply not available in a documented form. The initial selection had to be made on the best available knowledge of the trainers and the Regional Education Office staff.

Because of the difficulties involved in se ecting the sample we built an initial visit in January 1989 into the study in which the trainers were asked to collect data to determine the extent to which the above criteria were met. As a result the sample was cut down to twelve in the control and twelve in the experimental group. Two teachers subsequently left their posts and we thus finished up with a sample of 11 trained and 11 untrained teachers whose conditions are roughly comparable and on whom the study could be based.

A more careful screening of the schools as originally envisaged would obviously have been preferable. However, given the constraints in Nepal this was never feasible. In particular, the need to move the study forward at very short notice from its planned start date in March to January 1989 made these arrangements the best that could have been achieved.

Given the Nepali context and the nature of educational sampling in general we had no alternative than to base the study on an opportunity sampling. Random sampling was simply not feasible. We would have needed to select about three hundred teachers out of the total for upper secondary if this had been deemed necessary. This would have involved more time and expenditure than incurred in the rest of the project.

In summary, the study started with a selection of 16 control and 16 experimental schools which were considered likely to meet certain necessary conditions for inclusion. On the basis of screening visits in January 1989 8 of the schools which fell short of these criteria were removed from the study and we began the investigation with an n of 12 in each group. Since January, one trained and one untrained teacher left their schools and we eliminated their students' scores from the study.

Equivalence of the Groups

In the November 1989 visit we attempted to corroborate data on all the schools remaining in the sample in terms of: number of periods of language instruction received; continuation of teachers with the same group in year 9; additional training received by teachers in the untrained group; quality of pupil intake; overall academic performance as reflected by SLC results; student attrition.

Language Assessment of Teachers

During the November visit we were able to interview 18 of the 24 teachers and to assure ourselves that each had a base line language competence sufficient for them to teach the Nepali English curriculum in grades 8-10.

We managed to conduct tests on 18 of the 24 teachers in the study during the course of our visit. During the subsequent programme of visits the New Era staff administered all the tests with the exception of the oral to those not attending.

Method

Teachers (with the exception of the 6 not attending) were individually assessed on the basis of their performance in interviews, using the British Council's 9 band oral assessment checklist. To further determine their ability to teach English in the secondary system all teachers were given the students' tests which are based on grade 7 and 8 textbooks (the dictation and the gap filling). They were given an additional MCQ grammar test designed for University Entrance Language Proficiency screening in the UK.

In terms of their command of the structures and lexis in the books, there was little to choose between the two groups as was clearly shown in the dictation and gap filling tests. At this level they were both displaying a similar competence in the language. This is borne out by the t tests carried out on this data, where no significant difference can be shown between the control and experimental group teachers on the gap filling and dictation tests. Both groups exhibited a high degree of competence in these textbook based tests and their level contrasts sharply with the rather poor estimates of teachers' language ability highlighted in Davies' 1984 report.

In the analysis done on the students' test scores (DICT1/2, RDG1/2, WRIT1/2), teachers 'scores on which we had complete data were taken into account when assessing the students' improved performance from February to November.

In the statistical analysis we looked at the contribution of the teacher language level to student test performance and found there to be a negligible effect. There is no indication that teachers' language ability had any noticeable effect on students' language scores.

In terms of the student samples we have to accept the non equivalence of the two groups on the basis of their initial test scores but note that the differences are not large. In any case the General Linear Models Procedure (GLIM analysis) we used to analyse the data took these differences into account.

In terms of size of class, school results and the number of hours spent there are small differences between the control and experimental groups but these are not statistically significant . Size of class (SIZE), number of hours tuition (HOURS), School SLC results both general (GENPASS) and in English (ENGPASS) made insubstantial contributions to test scores.

5 Contracting The New Era for Data Collection

A decision was taken in November 1988 that local technical staff should be contracted to collect test and process data on the grounds of economy and their Nepali field experience. The New Era research organisation was selected as the best source for such staff.

5.1 Observer training January 1989

A second visit to the Nepal project was made by Weir to conduct an observer training session for New Era staff who would be responsible for collecting data on the effects of training on pedagogical practice. A special training manual was produced by Roberts and Weir for this purpose. In addition it was necessary to familiarise these staff in the conduct of the language tests. This involved a briefing on the instructions for invigilation and the steps to be taken post testing.

The observer training would seem to have been effective from what emerged in the trialling in Katmandhu. After joint observations the schedules were compared and a reasonable degree of agreement was noted. Where any differences occurred these were the subject of later training sessions.

The best four out of the six New Era staff (ie those who had performed best in the training) were selected to carry out the subsequent observations and the testing.

5.2 Monitoring visit November 1989

A further monitoring visit was made by Weir and Roberts in November 1989 with the following objectives:

a) To visit schools jointly with New Era staff.
b) To review language test procedures with New Era staff.
c) To review collection of observational data, particularly checklists.

d) To monitor the selection of sample schools and teachers.
e) To make recommendations for future data collection based on b-d above.

The following were the outcomes of the visit

re a) : Sixteen schools were visited, and thirteen teachers were jointly observed by New Era staff and Wir/Roberts between 5.11.89 and 10.11.89.

re b) : Language test procedures were reviewed in an initial briefing meeting with New Era staff in Pokhara on 4.11.89.

: The administration of tests was subsequently monitored in 5 schools and found to be satisfactory.

re c) : Observation procedures and category interpretations were reviewed and agreed in the meeting and a joint observation held on 4.11.89.
: Subsequent joint observations were reviewed and discussed.

On the basis of post lesson comparisons, there appeared to be a satisfactory level of reliability between observations made by Weir and Roberts and New Era staff.

6 Summary of the Data Available

By the end of 1989 the following data were available for analysis.

Language assessments

: Students' language tests: 22 schools (11 trained, 11 untrained); after removing outliers we had 716 students' script.
: Teachers' language tests: 22 completed tests; 18 oral estimates

Interviews

: 18 interviews (9 untrained, 9 trained)

Process Descriptions

:	Observations	:	22 teachers, 69 observation forms
:	Teacher self report	:	20 teachers (10+10), 54 reports
:	Sample student work	:	20 teachers (10+10)

6.1 Test Data

The analysis was carried out using SAS and in particular the General Linear Models Procdure (GLIM). As a first step the outliers in the population were removed from the sample by plotting the scores on graphs for each of the three tests, dictation, reading and writing . Their status as outliers was determined by their extreme position on the plotted scattergram. Candidates in the first administration of the tests scoring more than 15 on the dictation, or 24 on the gap filling, or 8 on the writing task, were removed from the sample as it was considered they were too dissimilar from the population we were interested in. This left us with an N of 343 students in the experimental group and 373 in the control group.

In all we had data on the performance of these two groups on the two sittings of the gap filling test (GAP 1 & GAP 2), the dictation (DICT 1 & DICT 2) and the writing (WRIT 1 and WRIT 2) In addition we were able to take into account the effect of a number of other variables on these test scores.

We had collected data on the teachers in the 2 groups on the same tests (DICT & READING) and on the grammar test (GRAM). The size of the classes attending the first test (SIZE) and the estimates of the number of hours of English each class had (HOURS) in the academic year 1989 were also available. The percentage of class time pupils spent talking in English (PUPENG) and the number of criterial features of training demonstrated by the teachers (FEATS) were also included in the analysis. Finally we have more limited data on the general pass rate of the schools in the SLC (GENPASS) and the English pass rate (ENGPASS) at the SLC.

6.2 Process Data

The observation schedule produces two quantifiable measures and supporting unquantified descriptions. The quantified daia consists of :

a) Pupil English : a raw number of pupil English (PE) codings, against codings for all kinds of talk, which can then be expressed as a percentage[PENG].

b) Criterial features : checklist entries which identify trained teachers'typical activities (cats.2,3,4,6,7,8,9,10,11)[FEATS] and untrained teachers' typical activities (cats 1,5).

Both these measures can be aggregated for the comparison of control and experimental groups. In our study, GLIM analysis included the variables of pupil English [PENG] and criterial features [FEATS]. The unquantified data in the observation schedules [notes with checklist entries and whole lesson descriptions] were used to conduct internal validity checks, by identifying the consistency between descriptions, checklist entries, and codings.

The 54 self report lesson descriptions were analysed by categorizing reported activities, identifying those associated with untrained and trained teachers, and displaying their relative incidence in the two groups.

Samples of student work were not analysed, as an insufficient number matched either the lessons observed or teachers'self reports.

7 METHODOLOGICAL ISSUES ARISING FROM THE STUDY

7.1 LANGUAGE TESTS

The use of the same test at the beginning and end of treatment is open to the criticism that any improvement may be due to practice effect. Given that our purpose is to compare the performance of the two groups we might reasonably assume that the practice effect benefits both groups equally. There was an eight month gap between the two administrations and students did not know they would be taking the same tests again. If we take scores on the first test into account in the analysis of the second administration this enables us to contrast gains made by the two groups. If there is a difference between the two groups in peformance on the second test administration it can be reasonally inferred that the training has had some effect. It would be imprudent, however, to try to isolate any specific training features as causes for observed changes in test scores. A cluster of associated variables result from training and it is not possible to identify with any certainty the relative contribution of individual variables to outcomes.

More difficult to answer are questions relating to the worth of any differences that might emerge. This is particularly the case in a non skills based test when one has to convert a quantitative score gain on discrete linguistic items into an interpretation. The judgement to be made on size of gain is in itself problematic when dealing with quantitative scores. If the gain is large the interpretation is better grounded.

In the case where the gain is relatively small one might wish to consider this from a longer term perspective. One might point to possible exponential as against linear gain in future test scores, given the possibility of initial inertia and old habits. If the teaching of English is to take place over a number of years any differences between groups might be magnified in future years. This of course argues that in studies of this type monitoring over a period of years would be required.

There is a critical need to ensure that some of the tests used in these studies are fair in content to both control and experimental group in order to make valid and fair comparisons. In addition there is a need to try to develop tests which are sensitive to particular features of training to confirm differential treatments. By definition the latter tests are not fair to both groups.

In situations such as Nepal where both groups are using the same course book and working towards the same final school examination, developing tests which were fair to both groups was possible by basing the test items on materials and activities common to both groups.

The greater problem lay in devising tests to reflect differences in pedagogical practice. We had to identify differences in treatment and develop tests which would measure effects on student performance. As we have indicated above this was problematic for two reasons. First, there is a general difficulty in establishing with any certainty causal relationships between pedagogical treatment and learning outcomes. Secondly, as we discuss below, there is a difficulty in identifying what the criterial features will be at the stage of the implementation of training rather than in the training itself.

There is a further problem where a study is designed to measure gain over a period of time in setting items at a suitable level of difficulty. If the items are too easy then a ceiling effect would quickly ensue and prevent any long term comparison. If the items are too difficult then it may be that they would be insensitive to gain even over an extended period. There needs to be a balance of items in terms of difficulty. We attempted to do this by basing the tests on Book 7 which all students had completed at the start of year 8 and also book 8 which they would have completed by the end of the first year of the baseline study. Given a very low start rate, poor previous learning experiences, and a limited number of hours of English, however, even items based on units covered may be beyond the reach of most of the students. In addition they may already be severely demotivated and this could interfere with the effects of any enhanced treatment.

It seemed sensible (in the absence of clear implicational items) to include tests with a large number and range of items in the first administration. In this way any differences would have a better chance of emerging and would enhance the reliability of the results obtained. There is always a danger that long tests might discourage students but observations did not suggest such an effect in this situation. It may be that statistical modelling through item response theory might be of some help in long term studies in the future.

It may well be that tests such as gap filling and dictation, because they focus on specific linguistic items may be testing constructs which take a long time to develop in learners. There is some suggestion in second language acquisition research that gains in linguistic competence may take a longer time to appear in comparison with skills development and performance. It may be the case that had we been able to develop practical tests of say spoken language ability, gains in test scores might have been clearly marked. This is an area which is in urgent need of research. The practical problems in testing skills such as spoken interaction cannot be ignored, however, and the limitations this imposes on evaluation studies are evident.

A final practical constraint is the length of time that the data even from a small scale study such as this takes to collect and process. At a conservative estimate it involved around 150 person days. This would be quite a sizeable chunk of a project member's time that would need to be allocated to evaluation.

7.2 PROCESS DATA

There are a number of reasons for treating the process data in this particular study with some caution and there are potential lessons for future evaluations to be learned.

1. The number of observations was not really sufficient to be sure of giving an adequate picture of teachers' customary practice. The conventional view is that about six visits are needed for this. Had New Era staff been able to obtain two observations per visit, the data would have been greatly improved. The requirement to complete tests within the shortest possible time, and in November, the imminence of the end of the school year, militated against this.

   This raises the question of the extent to which insiders should be involved, albeit in a formative evaluation role. For summative decisions the case for involving outsiders stands and it might be necessary to have a small number of outsider observations than none at all.

2. The reliability of observations is recognised as being greatly improved by the use of paired observers. For cost and logistical reasons this w·s not an option.

3. For reasons of extreme difficulty in access and travel, joint observations by New Era staff and the external evaluators were very much on an opportunity basis. In such circumstances one might not be able to conduct joint observations as one would want.

4. Quantification of process data is dependant on the identification of adequate units. In this study coding was based upon the recognition of utterance units rather than, for example, arbitrary time units. As a result coding boundaries were dependant on interpretation. For example when teachers or pupils repeat themselves, make false starts, or give one word responses, or during continuous speech such as in teacher explanation, it is possible that different observers might identify different numbers of utterances. In the case of question and answer exchanges, where counting speech utterance units is considerably easier, greater agreement can be expected. Ideally, it would be worthwhile to train observers in the the use of time unit coding, but considerable resources would have to be available.

5 The performance of observed teachers isoften influenced by "impression management". It may be that teachers provided "lessons to order" in which features of training would appear. The high occurrence of Pupil English in E group lessons and the high occurrence of Teachers' English rather than Nepali in C group lessons may over-represent the norm. Triangulation of data is the necessary strategy to validate observations. In this study a number of sources were used including teacher self report (see Appendix 2) and feedback forms. Our self report data suggested that criterial differences continued to be exhibited in unobserved lessons. The returns to an insider post course evaluation

questionnaire in 1988 presented a similar picture. Trained teachers reported some use of oral drills but in other areas, such as writing activities, they admit to extremely restricted application of training.

In spite of these limitations in the observational data, if the differences in criterial indicators between C and E groups are marked then they are likely to reflect real differences in classroom experiences for students.

7.3 LONG RANGE EVALUATION

Our experiences in Nepal highlighted the problems of outsider evaluation through long distance monitoring. As a rule evaluators working in this way are only able to spend a limited amount of time in the field. By necessity they have to work through others both in terms of setting up the study and in implementing it. Making practical arrangements such as meetings with teachers, visits to schools, organising transport and petrol, selecting a sample for the study, are all that much more difficult at a distance especially when internal communication in the country concerned is problematic.

In the absence of an extended field based feasibility study, outsiders also have to rely on insiders to provide them with information on which to construct the evaluation instruments. In our case we relied on the trainers to provide us with our information on the criterial features of training which they considered would be implemented by the experimental group. A feasibility study would have enabled us to scrutinise classroom practices as would attendance at a wider range of the training sessions. It would have enabled us for example to omit any concern with writing in Nepali classrooms from the study as subsequent experience demonstrated that because of the low importance of this in the school leaving examinations and because of time constraints on teachers, this activity was absent from most classroom practice. In terms of cost and time required, however, an extensive initial survey by outsiders may not be funded.

It does seem that this is a strong argument in favour of increased systematic internal monitoring by project staff. This would promote a more accurate definition of the categories of information for use in outsider evaluations. However, at crucial points outsider monitoring would be necessary as there is a risk of contamination of the data collected by personnel with an investment in the success of the project.

In our experience, contracting technical staff to conduct observations had unexpected benefits. We now realise that this means of collecting observational data is likely to produce an explicit analysis of criterial variables and the use of low inference criteria in observation along with less structured methods. As insider observers do not have the same need for expliciness or objectivity in producing observational data there is a real risk that their findings might not be meaningful to a wider audience.

7.4 ASPECTS OF SAMPLING

The results of our small scale survey are at best suggestive rather than conclusive as random sampling was not an option. In the event we were able to sample 11 out of the 1080 trained teachers.

We attempted as far as we were able to control for a variety of variables which might contaminate the results: class size, school leaving results, language level of teachers, attrition rates and number of hours of instruction received. By using the General Linear Models procedure to carry out the statistical analysis we were able to take account of these variables when determining the effect of treatment on language test scores.

Ethical problems arise in this type of non equivalent control group design. In particular there must be some concern about the anomalous position of the control group teachers. A small retainer was paid to encourage their participation in the study (preparing self report forms, attending meetings etc.). The effect of this has been to defer their training. Also we were acutely aware that in entering their classes our role was very much that of outsiders looking for evidence of deficit. Careful consideration needs to be given to their interests.

APPENDIX ONE

SEPELT BASELINE STUDY OBSERVATION FORM

OBSERVER	
TEACHER	
SCHOOL	
CLASS	Number of pupils present :
LESSON/ PAGE	
DATE	

<u>Teacher's Lesscn Outline</u>

APPENDIX A (Cont'd)

LESSON START □

TEACHER				PUPIL			
ENG		NEP		ENG		NEP	
Q	S	Q	S	Q	S	Q	S

START □

END □

ENG		NEP		ENG		NEP	
Q	S	Q	S	Q	S	Q	S

START □

END □

ENG		NEP		ENG		NEP	
Q	S	Q	S	Q	S	Q	S

START □

END □

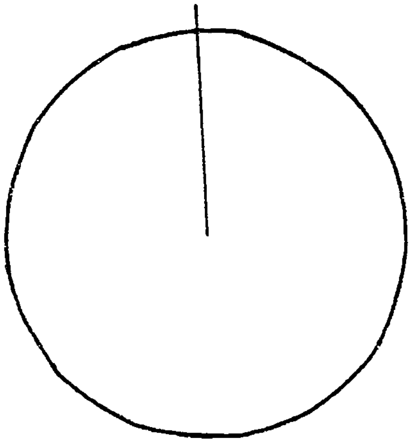
APPENDIX A (Cont'd) STAGES OF THE LESSON

from	to	
		1
		2
		3

	Teacher explained grammar mainly in Nepali		Pupils produced original sentences in English
	Teacher practised language in situations mainly in English		Pairwork or groupwork used
	Teacher gave models in English		Teacher gave listening practice
	Oral drills used,in English		Extra practice
	Pupils did written exercises by copying		
	Pupils did guided writing exercies		
	Comprehension questions asked in English		

Notes on extra practice:

APPENDIX A (Cont'd)



APPENDIX TWO

YOUR NAME Buddhi Prasad Sharma

LESSON DESCRIPTION

Please fill in the form, describing what happened in a recent, ordinary lesson: what you did, what the pupils did, the exercises you used, the work the children produced and so on. If you can, please attach an example of one pupil's work. Thank you.

Date May 15, 1989

Class VIII A
Sh. 69

WHAT I DID IN THE LESSON	WHAT THE PUPILS DID IN THE LESSON
- Teaching item: paragraph writing on 'How I make Haluwa' page 73	
- Eight instructional sentences from page 73 Exercise C were written on the card board before hand.	
- After reading the sentences, pictures and real objects were shown for difficult words.	
- Each sentence modeled twice e.g. Put a cupful of ghee in a pan. I put a cupful of ghee in a pan	Repeated by the pupils in chorus then individually.
- pupils were asked to write the same orally practised sentences omitting the numbers so as to make a paragraph	-- Class work was done by the pup
- while writing the paragraph, necessary check up of the sentences was made	- best paragraphs were read out by the pupils in the class
How. Some instructional sentences were given to write a paragraph on 'How I Make a Cup of Tea'	

PROGRAM EVALUATION IN LIGHT OF LANGUAGE LEARNING BACKGROUND, STUDENT ASSESSMENTS AND TOEFL PERFORMANCE

Harry L. Gradman and Edith Hanania

INTRODUCTION

We have carried out a study of the language learning background of students in our intensive English program to identify variables which have a positive effect on students' TOEFL performance. Data were collected from over 100 students in the program and the information coded and statistically analyzed. Using multiple regression and path analysis, we developed a model showing direct and indirect effects of a number of background variables on students' TOEFL scores. These variables, which include communicative use of English in class and extensive outside reading, suggest aspects of language teaching which may be used in program evaluation. The study also investigated the students' perceptions of their current language learning needs and their suggestions for improving language teaching in their home country. The students' assessments and the results of the statistical analysis provide useful guidelines for the evaluation of ESL programs at home and abroad.

At the Center for English Language Training at Indiana University, we have been interested in identifying factors in our students' language learning background which have a significant effect on language proficiency. Our main focus is students from other countries who have come to the United States to pursue higher education at American Universities. For the purpose of admission to academic university programs, the language proficiency test that is commonly used in the United States is the TOEFL (Test of English as a Foreign Language). We have therefore investigated background factors in relation to performance on the TOEFL examination, using as a reference point the initial TOEFL scores which the students obtained upon entering our Intensive English Program. We also looked into our students' perceptions of their language learning needs and their suggestions for improving language teaching in their home countries.

In this paper, we bring to bear findings from our research on the evaluation of ESL programs at home and abroad. We will first outline our procedure and summarize our findings, and then we will consider the implications of these findings for program evaluation.

PROCEDURE

The data for the study were collected by individual interview from 101 students in our Intensive English Program at Indiana University. The students came from a variety of first-language backgrounds, about equally distributed between Arabic, Japanese, Romance, and other languages. They had learned English in a formal environment to varving degrees, and their initial TOEFL scores with us were normally distributed. The background language learning information was elicited through an oral questionnaire and fell into four major categories: formal learning of English, exposure to and use of English in class, exposure to and use of English out of class, and attitudes and motivation. The items of information were coded and quantified, resulting in 44 background variables whose effects on the TOEFL scores were to be examined. Students' observations on aspects of their language learning background were also elicited and categorized.

The students' scores on TOEFL were used as a measure of language proficiency. TOEFL is a standardized instrument for language assessment which is widely used by universities in the United States for admission purposes. The examination is comprehensive and covers a range of language abilities, from elementary to advanced levels. Foreign students commonly must attain a specified minimum score as a prerequisite for admission to university work. The students in our Intensive English Program take an institutional version of the TOEFL at the end of each seven-week session. This test consists of three sections: Listening Comprehension, Structure and Written Expression, and Vocabulary and Reading Comprehension.

Three types of statistical analysis were used to examine the effects of background factors on language proficiency. The first was pairwise correlations between all the variables and the TOEFL scores, which led to the selection of a set of promising variables for closer examination. The second was multiple regression analysis, which further identified variables with a significant effect

on TOEFL. The third was path analysis on a basic structural model, which showed both direct and indirect effects.

RESULTS OF THE QUANTITATIVE ANALYSIS

The first type of analysis was pairwise correlations between all the variables. Table 1 shows the background variables which correlated significantly (p=.05) with TOEFL scores.

Table 1 Background variables which correlated significatly with the TOEFL scores (p=.05)

	r
Extracurricular (outside) reading	.53
Native speaking teachers	.39
English as language of instruction	.36
Months of intensive/special English	.26
Effective teachers	.21
Future need for English	.21

For a list of the variables, see Appendix 1

As can be seen from Table 1, the highest correlation coefficient was obtained for extracurricular reading. It may also be noted that the first four variables seem to reflect the extent of exposure to and use of English for meaningful communication. The last two are related to attitudinal and motivational factors. Surprisingly missing from this list are classroom variables, such as index of oral exposure, index of oral use, and index of writing use, as well as the extracurricular variables of writing, speaking and listening, all of which involve the use of English in communication. However, examination of the full intercorrelation matrices revealed that the above variables that did not correlate directly with TOEFL, nonetheless correlated significantly with outside reading, which correlated well with TOEFL. This suggested the presence of indirect effects, which were later explored through path analysis.

Based on the patterns of intercorrelation, 22 background variables were selected for further examination through multiple regression analysis. The results of this analysis, using the stepwise forward procedure (selection criterion p=.05), are given in Table 2.

Table 2 Multiple regression analysis for 22 background variables with the TOEFL scores

Steps		R	R2
1	Extracurricular reading	.53	.28
2	+ Native speaking teachers	.59	.35
3	+ Total index of exposure	.64	.40 (neg. beta)
4	+ Extracurricular speaking	.66	.43 (neg. beta)
All 22 variables (backward procedure)		.75	.57

For a list of the variables see Appendix 1

As can be seen in Table 2, the single most important background factor affecting performance on TOEFL is outside reading, followed by exposure to teachers who are native speakers of English. The two factors combined account for over one third of the variance in the TOEFL scores (R2=.35). The 22 variables together account for well over half this variance (R2=.57). The seemingly negative effects of total index of exposure and extracurricular speaking will in the next analysis (path analysis) be seen to conceal positive, though indirect, effects.

Path analysis was undertaken in order to explore the indirect relationships observed in the pairwise correlations and the unexpected negative effects observed in the multiple regression results. In this type of analysis, a causal model is formulated, consisting of a set of variables with interconnecting paths indicating the direction of effects. The statistical procedures calculate coefficients for the effects and evaluate the model by determining the extent to which it fits the available data. The statistical program we used for this purpose is LISREL (Analysis of Linear Structural Relations).

To date, we have constructed a basic model comprised of six variables: two classroom variables, three extracurricular (or outside class) variables, and the TOEFL scores. The paths lead from the classroom variables to the outside of class variables, and from these to outside reading and to TOEFL. The model showing the variables and the selected paths, all unidirectional, is illustrated in Figure 1. Preliminary results from the application of the statistical program LISREL to the full model indicated that a number of paths in the model did not represent statistically significant effects. The model was therefore trimmed to eliminate these relationships from the equation. The resulting modified model, showing the remaining paths and their coefficients, is presented in Figure 2.

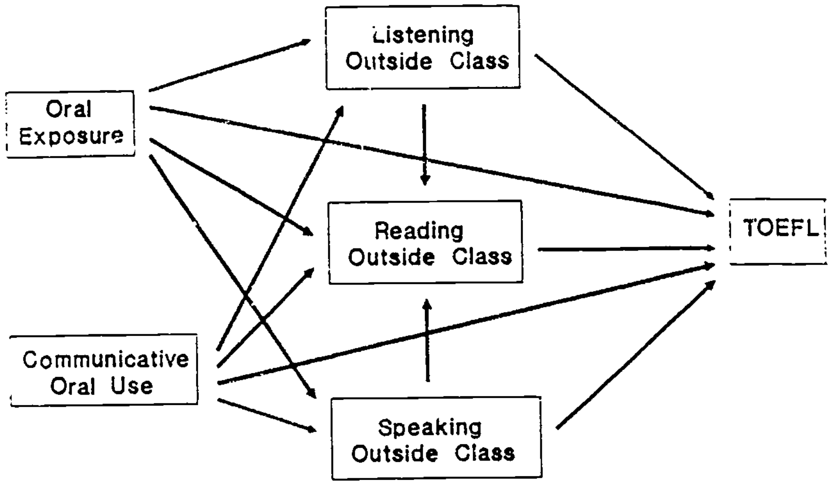


Figure 1

Structural Equation Model used for Path Analysis (LISREL)

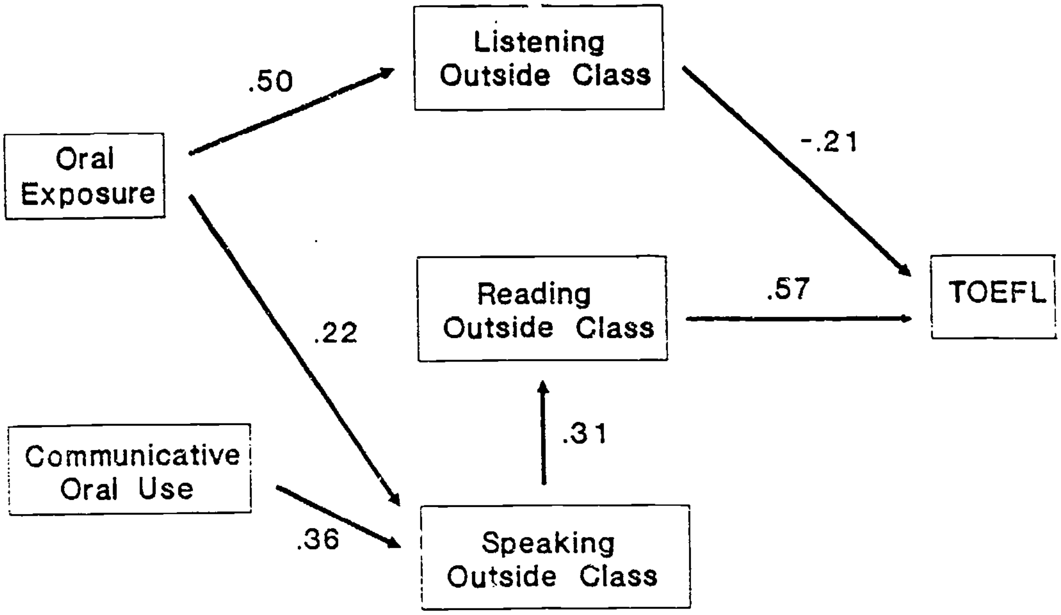


Figure 2

Modified Structural Equation Model Showing Significant Paths and their Coefficients

Goodness of fit index = .98

As can be seen, reading outside class again shows a strong direct effect on the TOEFL scores. The only other variable in this model that has a direct effect on TOEFL is listening outside class, but this relationship is a negative one. Speaking outside class, which had a negative relation with TOEFL in the multiple regression analysis, now shows a positive effect on TOEFL through its effect on reading. Likewise, the two classroom variables, oral exposure to English and communicative oral use of English, are now seen to affect TOEFL indirectly. Both have paths with positive direct effects on outside speaking, which in turn leads through outside reading to TOEFL. Oral exposure, in addition, has a path to outside listening, which has a negative relationship with TOEFL.

What this model seems to indicate is that the single most important factor in improving proficiency as reflected in the TOEFL scores is outside reading. Extent of oral exposure and communicative use of English in class and out of class have a positive, though indirect, effect on TOEFL in so far as they promote outside reading, but not through their effect on outside listening. Outside listening, which reflects the extent to which learners were exposed to English speech through radio, television, or film, appears to have a negative relationship with TOEFL.

Two aspects of these findings require further comment: the negative relationship between outside listening and TOEFL, and the strong positive relationship between outside reading and TOEFL.

Concerning outside listening, it should be noted that in our study the scores for this variable do not necessarily reflect active listening. In most cases, the students reported that, while viewing English programs, they relied totally on the native language subtitles and did not pay attention to or were unable to understand English speech. This type of exposure, therefore, did not represent comprehensible input, unlike the active listening included in the classroom variable, oral exposure. However, this explanation still leaves unaccounted for the negative relationship of outside listening with TOEFL scores. One possible interpretation is that students with higher outside listening scores may have a tendency towards passive viewing at the expense of more demanding activities, such as reading and social communication.

As for outside reading, the results of the study clearly appear to indicate that reading for personal information or pleasure is a more important means of implicit learning than exposure to and use of spoken English in and out of the classroom. It may be argued that the prominence of this factor reflects the degree to which good performance on the TOEFL depends on reading ability. While this may partly be the case, it seems unlikely that the use of reading as a medium for this examination can adequately account for the strength of the relationship of outside reading with TOEFL and its subsections. It would seem more likely that extensive outside reading helps to improve the level of proficiency in a global sense, enhancing acquisition of grammar, vocabulary, and rhetorical structure, as well as increasing the general knowledge base which helps in reading comprehension.[1]

In order to explore this point further, we ran multiple regression analysis using only those students in our sample who had entered at the three upper levels of our program (levels 4, 5, and 6) on the assumption that, for this more advanced group, basic reading skills had already been acquired. The results of this analysis (not tabulated) confirmed our finding for the whole sample. Outside reading was once again the most important, indeed the only, factor with a significant effect on TOEFL scores. Those students who read extensively out of class within this more advanced group attained higher levels of proficiency.

Another type of analysis that we applied to the data gave further support to the importance of outside reading in enhancing language proficiency. Using the t-test, we examined differences in the mean TOEFL scores of students who had done some outside reading and those who had not. We found that, even among students who had had the benefit of studying in an intensive English program, the mean scores of those who had done outside reading were significantly higher than those who had not.

The dominant role of outside reading which emerged from this study, although unexpected, is perhaps not surprising. Elley and Mangubhai (1983) in their experimental study on the effect of an extensive reading program on the language development of students in a number of Fijian primary schools found that students exposed to extensive reading of high-interest story books made significatly greater gains in language skills than the control group. In a more recent study, Tudor and Hafiz (1989) found that extensive reading improved students' writing significantly, particularly the level of accuracy. These findings are in accord with the theoretical viewpoint put forward by Krashen that reading, by providing extensive comprehensible input, is an important and effective means of acquiring language (Krashen 1981, 1988, 1989).

To sum up, the background factors which were found to have a direct positive effect on TOEFL scores were outside reading and teachers who were native speakers of English. Exposure to and use of spoken English in class and out of class for communicative purposes seem to be helpful only in so far as they promote outside reading. Although our present path analysis model does not include the variable native speaking teachers, it is reasonable to expect that a strong relationship holds between this factor and extent of exposure to and use of spoken English for communication in class and out of class. Indeed, pairwise correlation coefficients between these variables and native speaking teachers were quite high, ranging between .42 and .46.

RESULTS OF THE STUDENTS' OBSERVATIONS

Another aspect of the language learning background that we examined in this study was based on the students' qualitative observations. During the interview, we elicited from the students their comments on four aspects of their language learning experience: (1) why they liked or did not like their English class in high school; (2) what characteristics they valued in teachers of English whom they remembered as particularly good; (3) what they perceived to be their present language learning needs; (4) what their suggestions would be for improving the teaching of English in their home countries. Their responses, which were revealing, are outlined in this section.

Concerning the students' attitude towards learning English, about 45% of the students said that they had liked their English class when they were in school. The reasons they gave were fairly evenly distributed among the following categories:

(1) They thought English was useful or important. (27%)

(2) They liked English as a language, liked its sounds, liked to speak it. (20%)

(3) They regarded English class as an enjoyable period in which they could participate actively. (20%)

(4) They liked the teacher. (16%)

(5) They did well in English. (16%)

The reasons students gave for not liking English class represent the other side of the coin and are at least equally interesting. The reasons given, in order of frequency (high to low), are as follows:

(1) English was too difficult; they could not understand it and did poorly in it. (40%)

(2) The English class was boring, not interesting, a waste of time. (32%)

(3) They did not like the teacher. Reasons given included that the teacher was too mean, too strict, was not interested in teaching, could not speak or pronounce English well. (19%)

(4) English was not useful to them; it was imposed on them. (17%)

(5) The material taught was not relevant to their needs or interests. For example, they wanted to learn to speak but they were taught only grammar and reading. (6%)

These observations are echoed in the characteristics of teachers of English whom the students remembered as being particularly good. About 80% of the students said that they had had memorably good teachers. The characterizations that these students gave are listed below in the order of their frequency of mention:

(1) The teacher explained well, had a clear purpose, and was serious about helping the students learn English. (39%)

(2) The teacher had pleasing personality traits. Descriptions included: friendly, kind, encouraging, interesting, and having a sense of humor. (32%)

(3) The teacher used a variety of interesting activities and materials and encouraged the students to use English. (27%)

(4) The teacher spoke English well, had good pronunciation, and used English in class. (23%)

(5) The teacher made the students realize the importance of English in their lives and for their futures. (3%)

The next question concerned the students' perceptions of their current language needs. Of course, one must keep in mind that the students' responses reflect the fact that they are living in an English-speaking environment now and that most of them are planning to continue their education in American universities. It is therefore not surprising that the vast majority of the students felt that they needed more practice in listening and speaking. However, in addition, about 40% of the students mentioned the need to improve their reading and 6% their vocabulary through reading; about 28% recognized a need to improve their writing; about 9% felt they needed grammar, and about 4% said they needed to study hard. Two observations may be noted in connection with these results. The first is that our students' needs included all the language skills, singly or in various combinations, which confirms to us that none of these skills is superfluous in our program. The second is that there were important individual differences between the students, based on their past backgrounds and their future purposes, and that an intensive English program would do well to incorporate a measure of flexibility that would accommodate these differing individual needs.

In the last question, we asked our students to suggest how the teaching of English could be improved in their home country. Of course we realize -- and many of the students pointed this out -- that the teaching of English has changed dramatically over the past few years and that many improvements have been introduced. Nonetheless, the suggestions made by the students, based on their own experience and perceived needs, do provide valuable criteria for evaluating English language programs.

Most students called for the extensive use of English in English class and for increased attention to listening and speaking. Many stressed the importance of having teachers who are highly proficient in English, either native speakers or teachers who have received their professional education in English-speaking countries. Another set of suggestions relate to the interest factor in the classroom: varying the teaching methods and materials to include songs, games, tapes, movies, and reading analysis rather than mere translation. It was also suggested that reading and discussion should be on topics of current interest, and that the grammar should be of practical value rather than consisting of abstruse rules that are memorized with the object of passing exams. Several students also expressed the opinion that teachers should help their students recognize the importance of English in the modern world, encourage their efforts to learn the language, and make allowance for differences in individual ability within the class. We should perhaps add here that we were struck by the consistency in the observations the students made about their language learning experiences and their preferences, regardless of their native language and educational background.

To conclude, in our study we used the TOEFL examination scores of our students to help us identify aspects of their language learning backgrounds which contribute to language proficiency. We found that two background factors have a significant effect on TOEFL: extensive outside reading, and teachers who have an excellent command of English. We also found that communicative oral use of English in class and out of class affects performance on TOEFL through its positive effect on outside reading. These findings, along with the students' observations, which tend to highlight the importance of proficient and qualified teachers, provide a set of useful criteria for evaluating English teaching programs based, not on how students do on a particular test, but on the conditions that seem to promote the ability to function in the target language.

NOTES

[1] For the linguistic knowledge and complex processes involved in reading comprehension, see Goodman 1988, Carrell 1988, and Eskey 1988. For the relationship between general language competence and reading proficiency, see Elley 1984, and Devine 1988.

REFERENCES

Carrell, Patricia L. "Interactive Text Processing: Implications for ESL/Second Language Reading Classrooms." Carrell, Devine, and Eskey 239-59.

Carrell, Patricia L., Joanne Devine, and David Eskey, eds. 1988. Interactive Approaches to Second Langauge Reading. Cambridge: Cambridge UP.

Devine, Joanne. "The Relationship Between General Language Competence and Second Language Reading Proficiency. " Carrell, Devine, and Eskey 260-77.

Elley, Warwick B. 1984. "Exploring the ReadingDifficulties of Second-Language Learners in Fiji". Reading in a Foreign Language. Eds. J Alderson and A.H. Urquhart. London: Longman. 281-97.

Elley, Warwick B. & Francis Mangubhai. 1983. "The Impact of Reading on Second Language Learning". Reading Research Quarterly 19: 53-67.

Eskey, David E. "Holding In the Bottom: an Interactive Approach to Language Problems of Second Language Readers." Carrell, Devine, and Eskey 93-100.

Goodman, Kenneth. "The Reading Process." Carrell, Devine, and Eskey 11-21.

Krashen, Stephen D. 1981. Second Language Acquisition and Second Language Learning. Oxford: Pergamon.

Krashen, Stephen D. 1988. "Do We Learn to Read by Reading? The Relationship between Free Reading and Reading Ability". Linguistics in Context: Connecting Observation and Understanding. Ed. D Tannen. Norwood, New Jersey: Albex. 269-298.

Krashen, S D. 1989. "We Acquire Vocabulary and Spelling by Reading: Additional Evidence for the Input Hypothesis". The Modern Language Journal 73: 440-464.

Tudor, Ian and Fateh Hafiz. 1989. "Extensive Reading as a Means of Input to L2 Learning". Journal of Research in Reading 12: 164-78.

Appendix 1

Background Variables Considered

Formal Learning of English

*Age at start of English
Years of English in school
Years of English in University
*Months in intensive or special English programs
Months of private English tutoring
Contact hours for each of the above
*Total contact hours
Years since last English class
*Type of schooling: private/public

Exposure to and Use of English in Class

*Teachers who are native speakers of English
*English as language of instruction
*French or Spanish as language of instruction
Index of exposure to oral English
Exposure to instructions
Exposure to explanations
Index of oral use
Oral use - sentence practice
*Oral use communicatively
*Index of writing
Writing - sentence practice
Writing communicatively (composition)
Index of communicative use (oral + written)
Audio-visual index
Listening to tapes
*Language Lab
*Total index of exposure
(index of exposure + listening to tapes)
Teaching focus
(grammar translation/grammar reading/all skills)
Type of intensive/special English program
(home country/ES teachers/ES country)

Extracurricular Exposure to and Use of English

*Listening
*Reading
*Speaking
*Writing
Travel to English speaking countries

Attitude and Motivation

*Attitude to English class in high school
*Recognition of the need for English when in high school
*Effective teachers of English
*Family encouragement for learning English
*Family graduates from English speaking countries
*Current English use out of class
Current reason for intensive English
*Future need for English on return to the home country

* = the 22 variables selected for multiple regression analysis

TOWARDS EVALUATING THE WRITING LABORATORY: A PROTOTYPE

Ma. Flor E. Mejorada,
Elvira Fonacier

INTRODUCTION

Writing has been a major concern of researchers, language teachers and materials writers. A body of research has been done on the teaching of writing as a process (Flower 1980; Jacobs 1982; Zamel 1982, 1983; Raimes 1983, 1985;) which led to formulating appropriate strategies of responding to students' compositions; and eventually setting up scoring or rating techniques (Heaton 1975; Perkins 1983; Jacobs 1981 in Hartfiel et al. 1985; Josephson 1989). However, no one had so far come up with an evaluation scheme that includes various components of a writing program with which students are taught.

This paper proposes a model which attempts at a composite evaluation of the writing program. As such, it includes the learner's attitudes towards writing and the program, the progress he attains at the lab, and the delivery system of the program.

The proposal focuses on evaluating the Writing Laboratory (WL) of De La Salle University (DLSU) in Manila, Philippines.

A BRIEF DESCRIPTION OF THE PROGRAM

The Writing Laboratory of DLSU is a support unit of the Languages Department that accommodates students with writing problems. These students are recommended by their teachers for remediation. They take remedial classes for fifteen hours in the lab, spread over five weeks, meeting twice a week. They are taught by teachers who use the one-to-one conference-centered strategy. As there are only two to three students handled at one session the set-up makes immediate feedbacking possible. The WL's aims are to develop effective composing strategies and patterns of writing behaviour, to develop proofreading and editing skills and to facilitate the satisfactory writing and revision of academic papers.

SIGNIFICANCE OF THE MODEL

While the model reflects our perceptions of the Writing Laboratory, we hope to provide a framework which can be applicable to other writing contexts beyond the WL of DLSU.

The model speculates that the evaluation is primarily intended to help the program staff composed of teachers who actually conduct the lab and the program coordinator to examine the effectiveness of the program and of its identifiable subcomponents. The results will necessarily make them rethink assumptions that underlie the activities conducted in the lab. Secondarily, the evaluation will inform the administrators of the status of the program so that they can lend support in terms of deciding alternative courses of action to make the program serve best its target clientele.

SCOPE AND LIMITATION OF THE MODEL

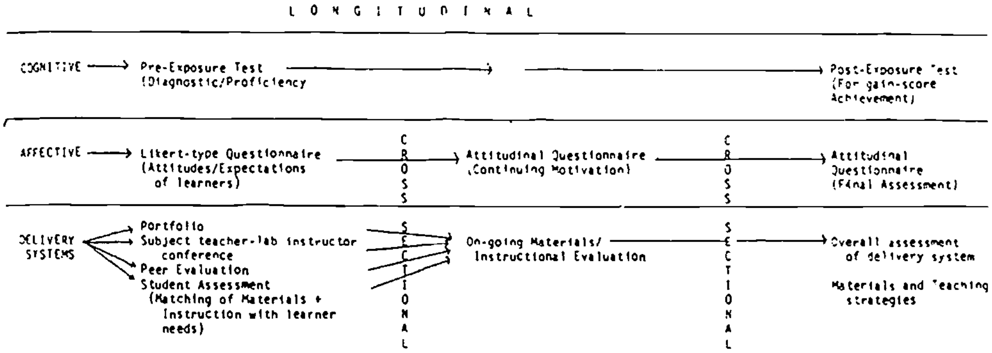
Viewed from the perspective of an on-going program delivery, the model covers three major components: affective evaluation, cognitive, and evaluation of delivery systems (Henning 1987). It is deemed that knowledge of the indexes of these three components could indicate the impact of the program on its users who are the students and its implementors who are the teachers as well as administrators.

To facilitate the modelling scheme, this prototype takes only twenty students whose attitudes toward the program and whatever skills they shall have attained will be presented. However, the model will not go into statistical details of the results of the questionnaires and tests given to them. Since the sampling was not representative of all the classes from where the lab population came, the model is not going to present an actual evaluation of the lab's achievement. Besides, cutting up the limited sample and analyzing against various components would only lead to finding something "significant" just by chance, and the small numbers would likely yield unstable interpretations. Therefore, considering these limitations, the model will only give a general view of the program's impact on the users.

METHODOLOGY

Respondents of the study. Through a pretest/diagnostic test, and a week's observation of class performance the teacher determines the students with writing problems and recommends them to the lab. However, enrollment at the lab is voluntary. So the students who actually register are presumed to have some degree of motivation which those of similar category do not have.

FIGURE 1 THE MODEL FOR EVALUATING THE WRITING LABORATORY


LONGITUDINAL
COGNITIVE
Pre-Exposure Test (Diagnostic/Proficiency
Post-Exposure Test (For gain-score Achievement)
AFFECTIVE
Likert-type Questionnaire (Attitudes/Expectations of learners)
CROSS
Attitudinal Questionnaire (Continuing Motivation)
CROSS
Attitudinal Questionnaire (Final Assessment)
DELIVERY SYSTEMS
Portfolio
Subject teacher-lab instructor conference
Peer Evaluation
Student Assessment (Matching of Materials + Instruction with learner needs)
SECTIONAL
On-going Materials/ Instructional Evaluation
SECTIONAL
Overall assessment of delivery system
Materials and Teaching strategies

The Design. Figure 1 illustrates the evaluation scheme which is both longitudinal and cross-sectional. It is longitudinal because it assesses the major components from the beginning through the last day in the lab; it is cross-sectional because it examines these components after the fifth meeting, that is the mid-part of the program. Here, the student's individual progress is considered, not for purposes of computing his grades but for correlating his written products with his realizations of the lab after getting acquainted with the materials and the strategies with which his/her instructor carries out activities. Additionally, these current discoveries will be interpreted in relation to records that are available in the portfolio.

Affective evaluation is carried out by means of a Likert-type questionnaire administered to students three times. the first one is to elicit feedback about their attitudes toward writing and to the program, and their expectations in the lab. The second and the last questionnaires are to know about their continuing motivation, which materials they find useful and which strategies they think work out for them. While the program may effect some attitudinal changes, the evaluation does not aim at statistically measuring them. It assumes that any positive attitudinal changes the learners have will be generally reflected in their writing performance. Besides, it is not customary to expect radical attitude gains in a short period.

Cognitive evaluation is determined through a pre/post test comparison. On the first meeting the student is shown the paper he wrote in class. His strong points and his weaknesses revealed in his composition are analyzed. He is then assured that his weaknesses can be remediated if he works hard together with his/her lab instructor. On the last day, he is asked to do a post test designed to be of equal difficulty to the pretest and the same test is also taken by those recommended by the teacher to the lab but chose not to enroll.

The posttest scores of the program takers also serve as an index to whether the instructional objectives were met or not. Questions on the appropriacy, measurability and attainability of the aims of the Writing Lab will be proved or disproved by the same results.

Evaluation of delivery systems is done through a portfolio monitoring done by the teacher, teacher ratings by students, and observations by fellow teachers and the WL Director, and a bi-weekly subject teacher-lab instructor conference. The teacher writes notes or keeps anecdotal records on each of her students as the instruction progresses. These notes are available to other teachers for purposes of comparing experiences and sharing information that are of common concern. The students' assessment is taken from their answers to question on strategies and materials in the attitudinal questionnaire. Although teacher ratings by students "have offered problems in that they have been founa to be more indicative of students' needs or desires than they are of teaching quality" (Henning 1987: 150), they are considered valuable in the lab to elicit reactions to the particular strategies not used in the regular classroom. Teacher observation done by the WL director is also of great value. The post observation conference is beneficial to both the teacher and the observer in terms of recognizing the program's strengths and scrutinizing problematic areas that call for new directions. Peer observation enables the instructor and fellow instructors to share with one another their on-going experiences with their respective students particularly on techniques and materials that have either proved successful or otherwise on certain writing problems of students.

Forty students with below 70% had similar ratings in the different components of the composition test they took on their first day in the regular English course.

Table 1: Pre Test Results of Students with Scores below 60%

Eng.Courses	N	Content	Organiz	Vocabry	LangUse	Mechnic	Total
ComArt31-Ex	3	17.00	14.00	14.33	16.00	3.67	65.00
ComArt31-Cg	5	15.60	14.00	14.40	15.00	3.60	62.60
ComArt41-Ex	6	15.67	15.50	14.67	15.33	3.83	65.00
ComArt41-Cg	4	16.00	15.25	15.25	14.50	4.00	65.00
EnglArt5-Ex	4	15.25	14.75	15.75	15.75	4.00	65.50
EnglArt5-Cg	4	15.50	16.50	15.50	14.50	4.25	66.25
EnComp12-Ex	5	16.60	16.20	15.00	15.40	3.80	67.00
EnComp12-Cg	3	14.33	15.00	16.67	16.67	4.33	67.00
EnComp14-Ex	2	15.00	16.00	17.00	16.50	4.00	68.50
EnComp14-Cg	4	14.50	15.00	16.00	15.25	4.00	64.75

AFFECTIVE EVALUATION

This evaluation considers the lab takers only. Since the focus is on what the program offers, the main aim for giving the questionnaire is to elicit feedback in order to give the maximum help to the students who have opted to be in the program.

IMPLEMENTING THE MODEL

RECOMMENDING STUDENTS TO THE LABORATORY

The regular classroom teachers, identified the students who needed the program. This was facilitated through the diagnostic test which they checked together with a lab instructor as interater. (Please see Appendix A, page 9, for a copy of the test). Moreover, the teachers' week-long observation of student performance in class would make them surer of who to recommend to the lab.

The scoring of the composition was based on Jacobs' ESL Composition Profile which includes the following components: Content (30%), Organization (20%), Vocabulary (20%), Language Use (25%) and Mechanics (5%). Because DLSU's passing cut-off score in the academic courses is 70%, those who got 69% and below on the diagnostic test were considered "failures" and therefore, needed the program.

Of the forty students recommended, twenty students actually took the course. They composed the experimental group (Ex) for the model. These students came from two sections of English for Computer Science (EnComp), section of English for Liberal Arts (EnglArt), and two sections of English for Commerce (ComArt2). The EnComp and EnglArt courses were under one teacher while the ComArt2 courses were handled by another. All of these English courses were basically writing with research along respective fields as focus. The rest of the poor writers, numbering twenty from the same classes, for whom the program should have been suitable opted not to take the program. These twenty students served as the control group (Cg). Table 1, p. 3 shows that.

On the first meeting with the lab instructor, the students answered a questionnaire (See Appendix B, .10). The answers indicated positive attitudes towards the program with majority claiming that the objectives of the lab were made known to them and that their coming to enroll was not because of any teacher pressure or classroom requirement. After the fifth meeting, these students had to answer another set of questions. This time the questionnaire was much longer since the purpose was to gain feedback about how the program was meeting the students' needs. (See Appendix C, pp. 12-13). On the last meeting, a questionnaire (Appendix D, p. 15) was answered again for final attitudinal assessment. Moreover, the results were expected to hint to the WL staff about what to keep, what to disregard and what to add to the program.

Table 2 below summarizes the results of the pre-laboratory, mid-laboratory and post-laboratory exposures survey of attitudes and continuing motivation of students. The means for the three surveys recorded values approximating 4.0 on the scale which could be interpreted as learners having positive attitude towards the program and having a rather stable degree of continuing motivation throughout the five-week lab period.

Table 2: Mean Scores on Attitude Survey

Score	Pre-Lab	Mid- Lab	Post-Lab	Mean Ave.
Mean	3.977	3.866	4.014	3.952

COGNITIVE EVALUATION

This component assessed the entry-level and exit-level performance of students (Please see Appendix E, p. 17 for the post test). Table 3 presents the post test results of WL students.

Table 3: Posttest Results of Students in the Writing Lab

Eng.Courses	N	Content	Organiz	Vocabry	LangUse	Mechnic	Total
ComArt31	3	21.67	16.00	15.67	21.33	4.33	79.00
ComArt41	6	18.33	16. 17	17.00	18.50	4.17	74.17
EngLart5	4	18.25	17.00	17.50	19.50	4.75	77.00
EnComp12	5	19.60	17.80	17.00	18.40	4.40	77.20
EnComp14	2	20.00	17.50	17.50	21.00	5.00	81.00

Figure 2, p. 6, presents a more detailed picture of students' writing skills noting their improvement in various components of the composition. Content (3.5 gain score) and Language use (3.7) recorded the first two highest gain scores. On the other hand, the scores of the Control group on the same components were much lower. Content recorded a 1.08 gain score and Language Use, 1.1.

Comparing the overall scores of the two groups, the lab takers registered a higher degree of improvement over those who did not take WL lessons. Figure 3, p.7 presents the picture of both the experimental and the control group on their pre/post performance with their respective gain scores. With the mean pretest scores obviously registering no significant difference between the two groups, the experimental group's posttest scores, which are higher than those of the non-lab takers (76.82 overall posttest score with a gain of 10.55 for the Ex group, and 69.58, with a gain of 4.23 for the C group), may permit the conclusion that their gain score can be attributed to their participation in the program. Of course, this does not discount the possibility that there might have been initial differences between the two groups which could not be identified, and that these differences could have accounted for some of the posttest differences.

FIGURE 2: PRE/POST TEST RESULTS

OF EXPERIMENTAL GROUP BY COMPONENTS

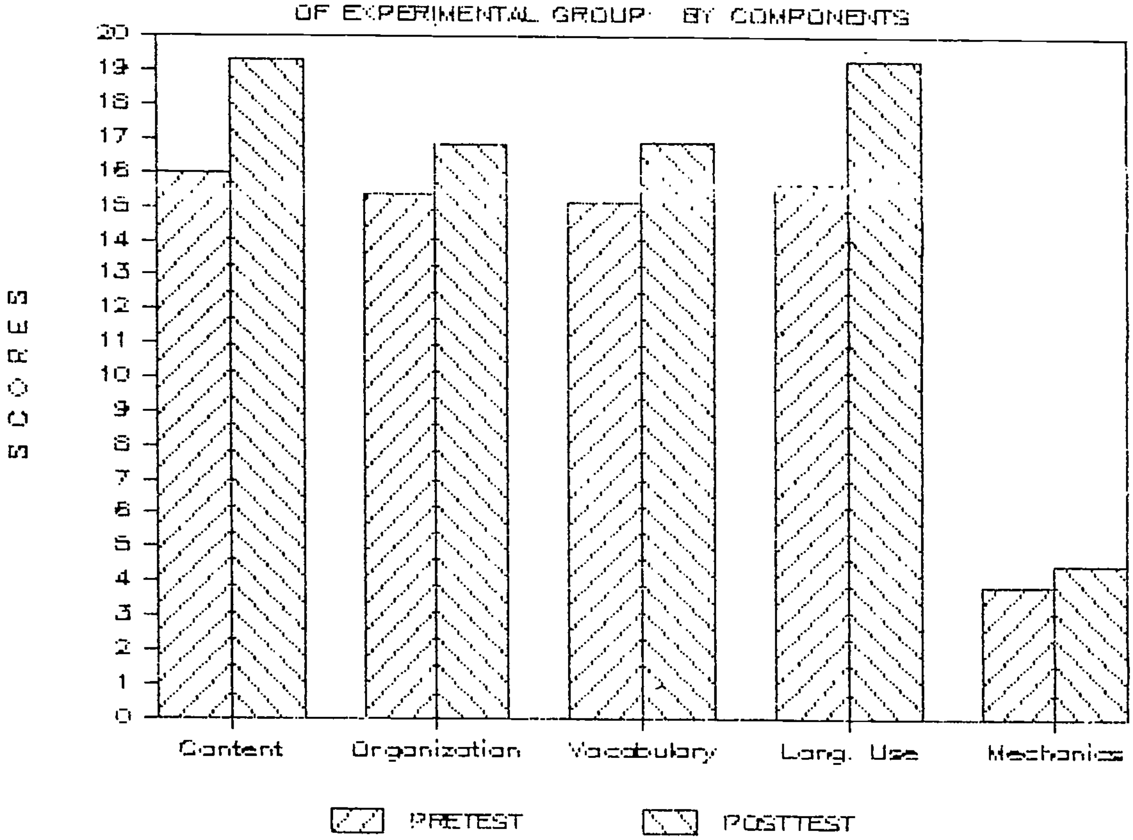
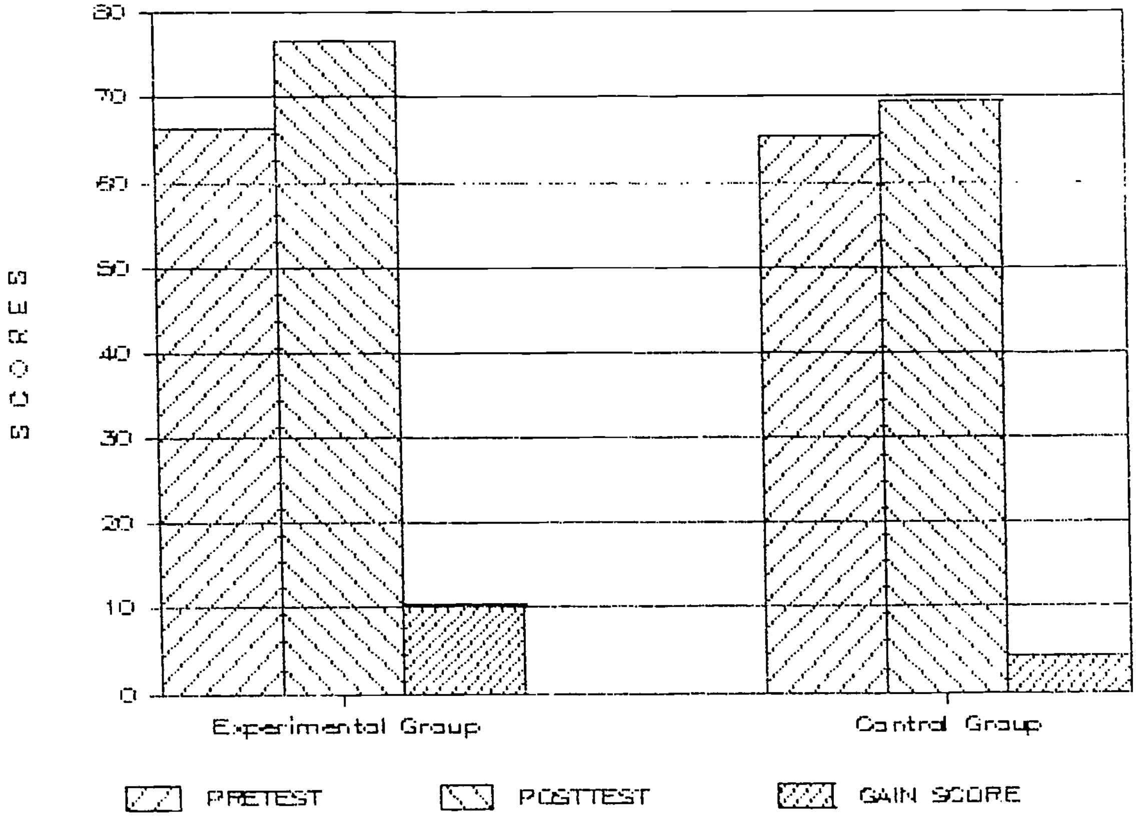


FIGURE 3: PRE/POST TEST RESULTS
OF EXPERIMENTAL AND CONTROL GROUPS



Henning, Grant. 1987. A guide to language testing: Development, evaluation, research. Cambridge: Newbury House Publishers.

Jacobs, Suzanne E. 1982. The writing of eleven pre-medical students. Washington D.C: Centre for Applied Linguistics.

Josephson, M I. 1989. "Marking" EFL compositions: A new method. English Teaching Forum. Vol. XXVII (3). 28-32.

Lewkowicz, Josephine A. Jayne Moon. 1985. Evaluation: A way of involving the learner. In J. Charles Alderson, ed. Evaluation. Oxford: Pergamon Institute of English.

Patton, Michael Quinn. 1982. Practical Evaluation. Beverly Hills, California: Sage Publications, Inc.

Perkins, Kyle. 1983. On the use of composition scoring techniques, objective measures, and objective tests to evaluate ESL writing ability. TESOL Quarterly. volume 17 (4). 651-671.

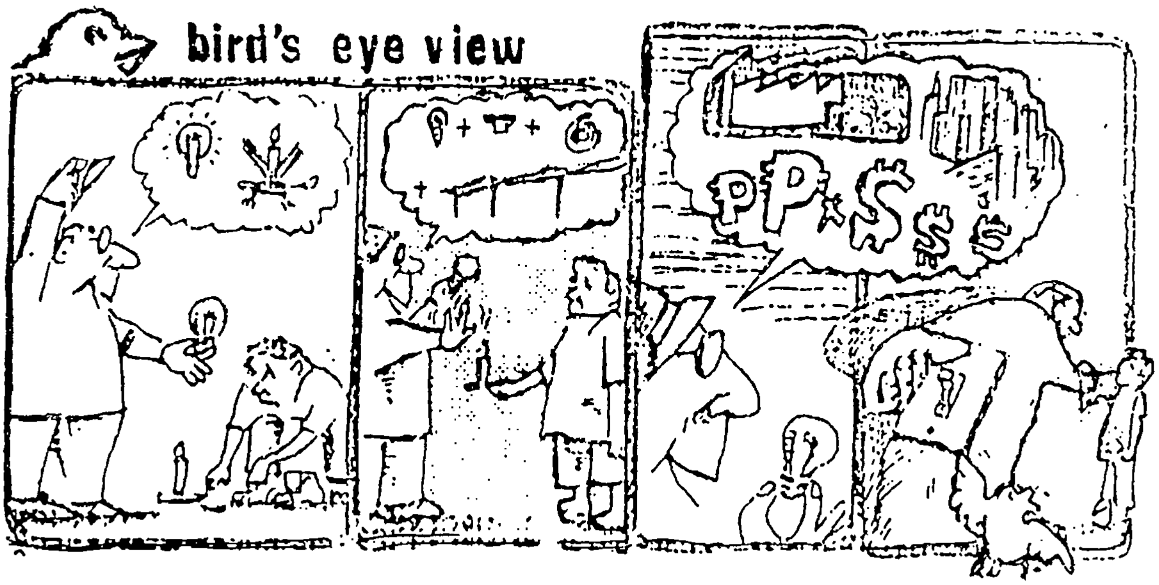
Raimes, A. 1985. 'What unskilled ESL students do as they write: a classroom study of composing'. TESOL Quarterly. 19: 229-58.

Zamel, Vivian. 1983. The composing processes of advanced ESL students: Six case studies. TESOL Quarterly 17 (2). 165-187.

APPENDIX A

Pre-Test

Direction: Study the picture below, and then write an essay of at least three paragraphs developing the topic, "The Fate of the Filipino Scientists Today" for the feature section of Bulletin Today (Daily newspaper).



APPENDIX B

Writing Laboratory Attitudinal Scale A

Instructions: Answer the following questions by shading the circle that corresponds to your answer in the sheet. The answer sheet has the following scale: 5 (Strongly Agree); 4 (Agree); 3 (Neutral); 2 (Disagree); and 1 (Strongly Disagree).

1 I expect to develop my writing skills in the laboratory.

2 The objectives of the program have been made clear to me.

3 I expect to identify my writing weaknesses.

4 I will attend the lab session regularly.

5 I am open to criticism regarding my writing.

6 I will make use of the lab resources to my advantage.

7 My going to the lab is of my own initiative.

8 I attend the lab to fulfill a requirement.

9 I enrolled in the lab because of teacher-pressure.

10 I do not feel comfortable with my writing.

APPENDIX B (Cont'd)

Teacher : ______________________________ Section: ______

Form A

	5	4	3	2	1
1.	○	○	○	○	○
2.	○	○	○	○	○
3.	○	○	○	○	○
4.	○	○	○	○	○
5.	○	○	○	○	○
6.	○	○	○	○	○
7.	○	○	○	○	○
8.	○	○	○	○	○
9.	○	○	○	○	○
10.	○	○	○	○	○

APPENDIX C

Writing Laboratory Attitudinal Scale B

Instructions: Answer the following questions by shading the circle that corresponds to your answer in the answer sheet. The answer sheet has the following scale: 5 (Strongly Agree); 4 (Agree); 3 (Neutral); 2 (Disagree); and 1 (Strongly Disagree).

1 The materials are helpful in attaining the objectives of the program.

2 The instructor helps me improve my writing abilities.

3 I find the program addressing my writing weaknesses.

4 I feel that my writing skills have improved in the lab.

5 The time I spend in the lab is worthwhile.

6 I find the length of time I spend in the lab sufficient.

7 The one-on-one instruction is helpful in my achieving the objectives of the course.

8 The quality of materials meets my expectation.

9 The instructor explains well and to the point.

10 My writing has improved since I attended the lab.

11 I am getting my money's worth in the lab.

12 My writing had developed significantly with the program.

13 The self-access materials were sufficient to address my needs.

14 I have developed confidence in my writing abilities.

15 I have learned to evaluate my own writing and revise it according'y.

16 I find the instruction in the lab adequate.

17 I find my going in the lab a waste of time.

18 There are materials which I find difficult or irrelevant.

19 The individualized instruction is not helping me at all.

20 The 15-hour session in the lab is too short to enable me to develop my writing skills.

21 I feel that the one-and-a-half hour sessions should be longer.

22 My instructor is most of the time unprepared for the session.

23 Choosing materials in the self-access collection is difficult.

24 The lab needs to develop better materials for instruction.

25 There is a mismatch between my needs and what the lab has to offer.

26 The objectives of the lab were not clear to me.

27 I am not able to develop my writing skills despite the program offerings.

28 My instructor does not provide enough feedback to allow me to see my progress.

29 The lab program needs further development.

30 I fell that I am getting poor quality instruction in the lab.

31 My writing has remained the same as when I started in the lab.

32 I don't feel that the lab is able to address my writing problems.

33 I always feel threatened everytime my instructor comments on my composition.

34 My morale has been very low since I attended the lab sessions.

35 I feel at ease working in the lab.

36 I feel at ease conferring with my instructor in the lab.

37 I am bored with the way my instructor conducts the sessions most of the time.

38 The lessons are repetitive of those I have in my English course.

39 The lessons in the lab complement those in my English course.

Teacher: ______________________ Section: ________

Form B

	5	4	3	2	1		5	4	3	2	1
1.	0	0	0	0	0	21.	0	0	0	0	0
2.	0	0	0	0	0	22.	0	0	0	0	0
3.	0	0	0	0	0	23.	0	0	0	0	0
4.	0	0	0	0	0	24.	0	0	0	0	0
5.	0	0	0	0	0	25.	0	0	0	0	0
6.	0	0	0	0	0	26.	0	0	0	0	0
7.	0	0	0	0	0	27.	0	0	0	0	0
8.	0	0	0	0	0	28.	0	0	0	0	0
9.	0	0	0	0	0	29.	0	0	0	0	0
10.	0	0	0	0	0	30.	0	0	0	0	0
11.	0	0	0	0	0	31.	0	0	0	0	0
12.	0	0	0	0	0	32.	0	0	0	0	0
13.	0	0	0	0	0	33.	0	0	0	0	0
14.	0	0	0	0	0	34.	0	0	0	0	0
15.	0	0	0	0	0	35.	0	0	0	0	0
16.	0	0	0	0	0	36.	0	0	0	0	0
17.	0	0	0	0	0	37.	0	0	0	0	0
18.	0	0	0	0	0	38.	0	0	0	0	0
19.	0	0	0	0	0	39.	0	0	0	0	0
20.	0	0	0	0	0						

APPENDIX D

Writing Laboratory Attitudinal Scale C

Instructions: Answer the following questions by shading the circle that corresponds to your answer in the answer sheet. The answer sheet has the following scale: 5 (Strongly Agree); 4 (Agree); 3 (Neutral); 2 (Disagree); and 1 (Strongly Disagree).

1 With the program, I can now write a well-organized composition.

2 I feel that I wasted my money in the lab.

3 The materials to work on were those that I did not need.

4 The lab program is an unnecessary support of the English course.

5 I am confident of what I am doing in the lab most of the time.

6 I feel that I am not using my time well in the lab.

7 Until now, I could not write a composition that meets the standard of my instructor.

8 My instructor has been very supportive throughout the program.

9 I feel I have accomplished very little in the lab.

10 The lab sessions have been productive.

Teacher: ______________________________ Section: ______

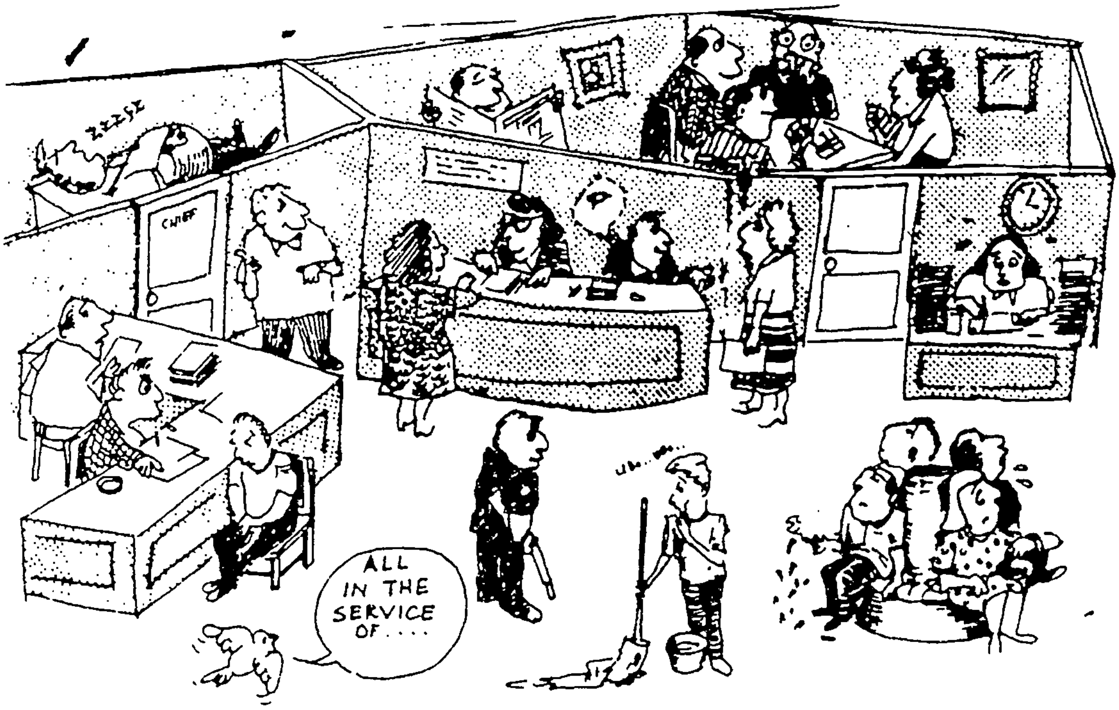
Form C

	5	4	3	2	1
1.	0	0	0	0	0
2.	0	0	0	0	0
3.	0	0	0	0	0
4.	0	0	0	0	0
5.	0	0	0	0	0
6.	0	0	0	0	0
7.	0	0	0	0	0
8.	0	0	0	0	0
9.	0	0	0	0	0
10.	0	0	0	0	0

APPENDIX E

Final Composition

Direction: A local newspaper is inviting you to write an editorial for its next issue. Based on the picture below, write an essay of at least three paragraphs answering the question, "What Makes an Employee Unproductive?"



TRIALLING OF THE NEW EL SYLLABUSES FOR SINGAPORE SCHOOLS

Goh Soon Guan

INTRODUCTION

1 New English Language syllabuses have been prepared by the Ministry of Education and will be implemented, in stages, in all primary and secondary schools in Singapore beginning in 1992. In this paper, I will report on the trialling of the English Language syllabuses for the lower blocks, ie the lower primary and lower secondary classes. [In Singapore's education system, lower primary refers to the first 3 years of formal education in the elementary school, while lower secondary refers to the first 2 years of the secondary or junior high school]. These new syllabuses have already been distributed to all schools. The full syllabuses for both lower and upper blocks will be distributed to schools in early 1991.

OUTLINE OF PAPER

2 In Part I of this paper, I shall set out very briefly the rationale for revising the existing syllabuses in schools and then describe, again very briefly, the organisational framework of the new syllabuses. In Part II, I shall mention the objectives and then talk about how the trialling exercise was carried out. Finally, I shall share with you the findings of the trialling, which were captured mainly through a questionnaire completed by the participating teachers. These findings were confirmed through observations of lessons in the classrooms and through informal discussions with the teachers.

PART I - RATIONALE FOR REVISION

3 The existing English Language syllabuses were published in the early 1980s in conjunction with the New Education System in which streaming was first introduced in both primary and secondary schools. These syllabuses stipulate the amount of language learning at each year level, and contain a list of grammar items as well as skills for developing listening, speaking, reading and writing competence. These syllabuses are rather examination-oriented, and although certain principles underlying teaching methods are given, no particular approach for teaching English Language is recommended to the teacher.

RATIONALE BEHIND THE NEW SYLLABUSES

4 The new English Language syllabuses take a more comprehensive approach to language teaching and learning. Modern approaches to language teaching and learning, such as communicative language teaching, learner-centred pedagogy and process-oriented methods in reading and writing are taken into account. Common to these approaches are the emphases on the processes of language learning, the strategies and techniques used by good language learners, and the interaction and learning which result from different forms of classroom organisation and activities. These approaches are reflected in a number of innovative English Language programmes and projects which were implemented in our schools in the last five years, and which have been found to be effective in helping our pupils to acquire their language skills more quickly and even to enjoy language learning. I shall refer, in particular, to 3 of these programmes. They are the REAP and ACT programmes in our primary schools, and the PASSES project in selected secondary schools. REAP, which stands for Reading and English Acquisition Programme, is an interesting programme in which lower primary pupils learn English through lots of reading and language experience. ACT, which stands for Active Communicative Teaching, is an activity-based programme for upper primary pupils and stresses the integration of the four language skills. And PASSES,

which stands for Project to Assist Selected Schools in English Skills, represents the most effective strategies to help weaker schools in upgrading their English Language programme. The basic principles of effective language teaching and learning, inherent in these programmes, are now reflected in the new English Language syllabuses and will be features of all English lessons when the syllabuses are implemented.

5 Another important consideration when revising the existing syllabuses was the importance of taking advantage of the information explosion and the advanced technology seen all around us to support our English Language programmes. The new syllabuses encourage teachers to exploit fully the benefits of extensive reading programmes as well as the power of information technology, particularly television, radio and tape, and to use more extensively other informal ways of learning such as dramatization, project work and collaborative learning. Also, to train pupils to cope with the vast amount of information and to equip them with life skills for more independent learning, the new syllabuses specify skills on learning how to learn, eg study and information skills; thinking skills; and even learning language through a better understanding and appreciation of culture.

6 In order to help our pupils build on their language skills more quickly, the new syllabuses, unlike the old ones, do not specify the amount of language learning at each year level. Instead, terminal objectives are specified for each primary or secondary block. These terminal objectives are grouped under 4 domains, viz

A Communication and Language Development
B Thinking Skills
C Learning how to learn
D Language and Culture

The provision of terminal objectives for each block allows pupils to learn much earlier certain language items or skills at a higher level within the block if they are more proficient in their language, and then proceed to the upper block and a more challenging set of terminal objectives.

ORGANISATIONAL FRAMEWORK

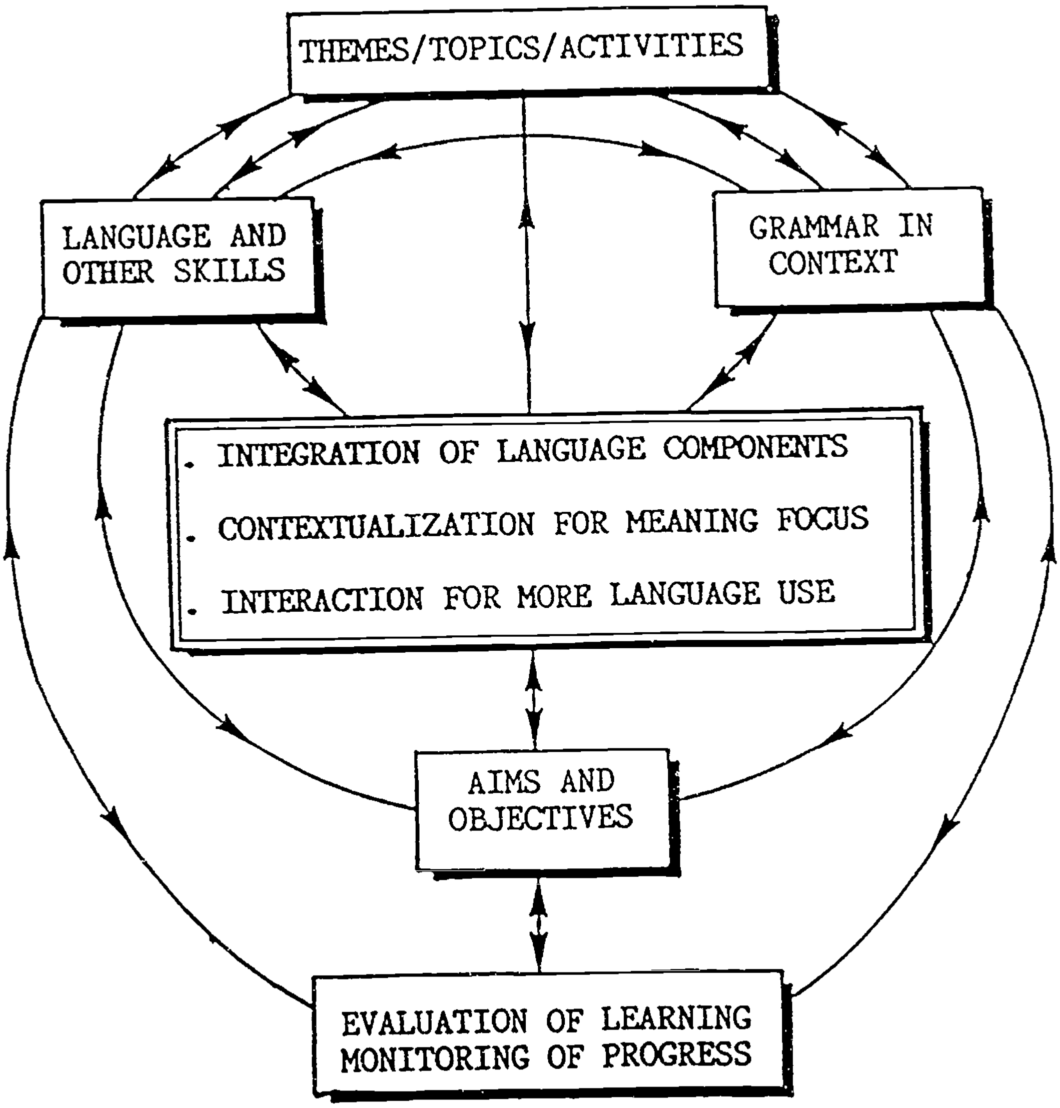
7 Since the new syllabuses stress the processes of learning and the procedures of teaching ie, methodology, we have formulated an organisational framework in which all the components are integrated to reinforce and maximise learning. This organisational framework is best explained with the use of this diagram (see Fig 1 attached). Central to this framework is the need to contextualize learning and to integrate the various skills and components in order to facilitate learning. This contextualization and integration is achieved through the use of themes which cover a wide range of topics to cater to the varied interests and maturity levels of the pupils as they progress through their school years. A list of these suggested themes can be found in our new syllabuses.

8 The themes provide the context for the teaching and learning of language and communication skills, as well as grammar, through meaningful activities. These activities are planned with appropriate objectives in mind. Through interaction in such activities (which are supported by the use of various resources) pupils develop the relevant skills as set out in the terminal objectives. These activities also provide the means for integration of the various language components. It is this integration, together with the interesting tasks and activities, which makes language learning more purposeful, more meaningful and more motivating for our pupils.

9 Finally, by monitoring and evaluating this pupil-centred learning process, the teacher gets feedback on his pupils' progress. This feedback is essential for the teacher to plan sufficiently challenging language tasks to sustain his pupils' interest and motivation in language learning.

10 The rationale and content of these six key elements (themes, skills, grammar, integration, objectives and evaluation) are elaborated in the six chapters in the syllabuses :

Figure 1 The inter-relationships between the various components of the syllabus


THEMES/TOPICS/ACTIVITIES
LANGUAGE AND OTHER SKILLS
GRAMMAR IN CONTEXT
. INTEGRATION OF LANGUAGE COMPONENTS
. CONTEXTUALIZATION FOR MEANING FOCUS
. INTERACTION FOR MORE LANGUAGE USE
AIMS AND OBJECTIVES
EVALUATION OF LEARNING
MONITORING OF PROGRESS

Chapter 1 sets out the objectives of learning English based on the various needs of our pupils.

Chapter 2 discusses the pedagogic approaches based on theoretical principles that are drawn from various theories about language and learning. The integrated approach that is advocated is exemplified in a sample sequence of lessons.

Chapter 3 consists of an inventory of suggested themes and topics as well as an inventory of tasks and activities for different areas of language learning.

Chapter 4 gives the spectrum of skills, a list of micro skills for language learning, encompassing the different language components, including thinking skills, learning how to learn skills and skills relating to language and culture.

Chapter 5 is an inventory of grammar items which teachers could consult for their planning of tasks and activities.

Chapter 6 provides some guidelines for assessment and evaluation of language learning to help teachers assess their pupils' learning progress fairly and reliably. Ideas on self assessment and pupil profiling are included.

PART II - OBJECTIVES OF TRIALLING

11 Plans were made in 1988 for trialling the draft syllabuses in a representative sample of schools. The objectives of the syllabus trialling exercise were :

- To find out if teachers can understand and interpret the contents of the syllabus
- To find out if teachers can plan a framework for an integrated sequence of lessons* (about 2 weeks) by making effective use of the various inventories in the syllabus
- To find out if teachers can select and adapt relevant materials for teaching and learning, based on a theme chosen by themselves
- To find out if teachers encounter problems in implementing the planned sequence of lessons in their classes
- To obtain the above feedback from the teachers for revising the syllabuses and for planning training workshops for teachers in using the syllabuses

* An integrated sequence of lessons has the following main features :

- It consists of a series of lessons linked by a theme or topic
- It is carried out over two weeks or more
- It does not compartmentalize the language components
- It combines the learning of all language and language-related skills, functions, grammar and vocabulary
- It builds in monitoring and evaluation of language learning

METHOD USED

12 A stratified random sampling technique was used in the trialling exercise. 12 schools were identified based on their consistently high, average or weak results in English Language at the two public examinations, the Primary School Leaving Examination and the GCE 'O' Level Examination. There were six schools from the primary section, involving three teachers in each school teaching Primary 1, 2 and 3 levels, and six schools from the

secondary section involving two teachers in each school teaching Secondary 1 and 2 (either Express or Normal streams. [In the Express stream, pupils take the GCE 'O' Level Examination after four years of study, while in the Normal stream, pupils take this examination in their fifth year, if they perform well in the 'N' Level Examination]. In all, 30 teachers from 12 schools located in different parts of the country participated in this exercise.

13 In early 1989 teachers from the 12 schools selected for trialling received photostat copies of the relevant draft syllabuses for reading. Two workshops were then conducted for the teachers on interpreting and using the draft EL syllabuses, including how to plan an integrated sequence of lessons. The teachers then planned their actual sequence of lessons to be carried out in their own classes. They were assisted by a Specialist Inspector attached to each school, who also visited the classroom to see the lessons, to share ideas for teaching, and to team-teach if necessary. Finally, the teachers completed a questionnaire on the trialling of the draft English Language syllabuses.

FINDINGS

14 Part I of the questionnaire requested for a profile of each teacher. Feedback indicated that all the teachers were qualified and trained and some had additional teaching qualifications. Most of the teachers had been teaching for 5 to 15 years.

In Part II of the questionnaire, the teachers were asked to respond to 29 statements using a Likert-type scale anchored by Strongly Agree, Agree, Uncertain, Disagree and Strongly Disagree. The 29 statements were listed under 4 headings viz

A The Draft Syllabus as a Document
B Planning an Integrated Sequence of Lessons
C Implementation of the Integrated Sequence of Lessons
D Teacher Training and Materials Development

A The Draft Syllabus as a Document (Please refer to Appendix, Statements A1 to A6)

15 It was heartening to note that more than 70% of the teachers considered the chapters in the syllabus well organized and that 66.7% of the teachers found cross-referencing between chapters easy. On whether the syllabus had an overload of information, 50% from the primary section and 36.4% from the secondary section agreed. A considerable proportion of the teachers (about 25%) was uncertain.

16 Many teachers differed in their opinion about the concepts in the syllabus being not well explained, although more teachers disagreed with this statement. This negative response could be attributed to the view in Statement 5 (subscribed to by most teachers) that the language used was too technical. However, some comments given by the teachers in the questionnaire as well as in discussions during the two workshops prior to the trialling exercise indicated that it was the unfamiliarity with some linguistic terms and concepts used in the syllabus, as well as the formal register, that caused the difficulty in understanding and hence the negative response.

17 Finally, on the statement whether the syllabus provided clear guidelines for assessing pupils' progress, the responses were rather negative from the primary section but less certain from the secondary section. Teachers' unfamiliarity with certain concepts on assessment and some of the terms used, leading to minimal understanding, together with the fact that this chapter in the syllabus was not used in the planning of the integrated sequence of lessons, were possible explanations for the kinds of responses indicated.

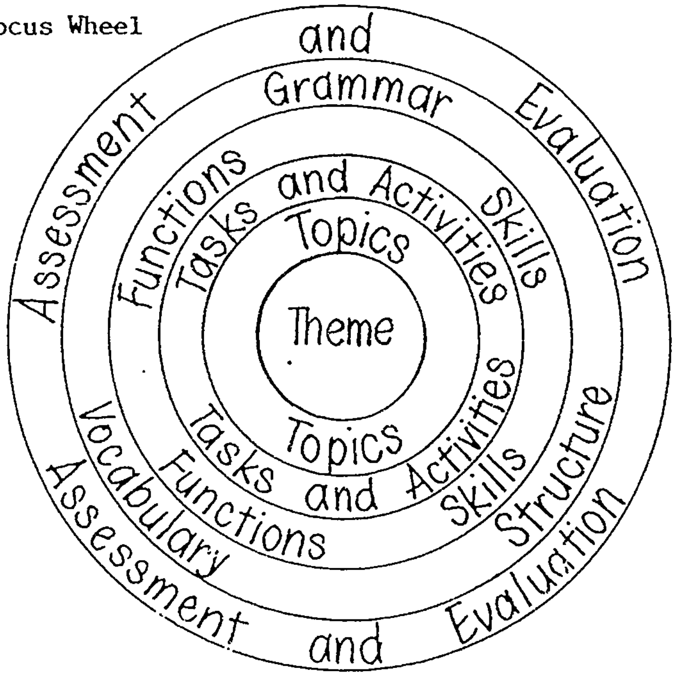
18 The findings for this section of the questionnaire were expected. It must be mentioned that many of the teachers had stated that they were not used to reading such a large quantity of material in a syllabus. (The new syllabuses contain twice as much material as the old syllabus). Moreover, many of them had only read the whole document once or

only referred to the chapters or sections within chapters that were needed for the planning of the integrated sequence of lessons.

B Planning an Integrated Sequence of Lessons (Please refer to Statements B1 to B11)

19 There was unanimous agreement in the primary section (100%) but less so in the secondary section (83.3%) that the syllabus could be used flexibly for lesson planning and for planning an integrated sequence of lessons. Also, the majority of the teachers found the examples and the Focus Wheel in Chapter 2 helpful in lesson planning. [The Focus Wheel (see Figure 2 below) is a concept used in planning an integrated sequence of lessons. It consists of a series of concentric circles. Starting from the centre of the circles is the theme followed by the topics, from which the teacher plans the tasks and activities. The teacher then thinks of the skills and functions, and the vocabulary and language structures that can be generated from the tasks and activities. Finally, ideas on assessing or evaluating what pupils have learnt are built in at the end of the sequence of lessons].

Figure 2. A Focus Wheel



20 The teachers viewed the suggested approach to lesson planning very positively. They agreed that the suggested approach encouraged the use of a variety of resources and teaching strategies and also encouraged meaningful language use among the pupils. However, the majority of teachers (above 90%) felt that this type of planning took up too much time. The teachers' response here could be linked to their agreement with Statem... 11 concerning the difficulty of obtaining suitable materials for teaching and lear...g. Discussions with many teachers during the trialling exercise confirmed that teachers spent a great deal of their time searching for suitable materials for the topics on which they were planning their English Language lessons.

21 On Statement 9, pertaining to the possibility of covering an adequate range of language items (eg grammar, functions, vocabulary, skills), only 50% of the teachers were in agreement. About 40% of the teachers did not think that it was possible. This fairly large proportion of negative responses could be related to their response in Statement 10, where the majority of teachers agreed that the instructional objectives were not easily derived from the tasks and activities. We suspected, however, that many teachers did not know how to exploit a task/activity for particular or more varied instructional objectives. This was found out to be true during visits to teachers to assist them in the planning of an integrated sequence of lessons, when many teachers found difficulty in matching instructional objectives with tasks that they had planned for pupils.

Assessment of Inventories in the Syllabus (Please refer to Statement 12)

In this sub-section, teachers were requested to rate the various inventories listed on the left according to whether each inventory was comprehensive, useful and easy to use.

Themes and Topics

22 There was strong support that the themes in the syllabus were comprehensive, useful and easy to use.

Tasks and Activities

23 Most teachers considered the tasks/activities inventory very comprehensive and useful but not so easy to use (only 50% agreed in the primary section).

Grammar

24 About 50% of the primary teachers thought that the grammar inventory was comprehensive and useful but not so easy to use (58.8%). On the other hand most secondary teachers thought that the grammar inventory was comprehensive and easy to use but were uncertain about its usefulness (only 50% positive).

Communicative Functions

25 Responses from the primary teachers that this inventory was comprehensive, useful and easy to use were slightly more than 50%. The secondary teachers thought that this inventory was comprehensive but did not consider it very useful or easy to use.

Spectrum of Skills

26 Most primary teachers did not consider this inventory very comprehensive while secondary teachers thought that this inventory was sufficiently comprehensive but not very useful or easy to use.

27 Overall, the teachers were generally positive about the inventories of themes, topics, and tasks/activities but they did not consider the inventories of grammar items and skills very favourably. One possible reason was that many teachers were uncertain about how to teach grammar and skills in a more communicative way.

C **Implementation of an Integrated Sequence of Lessons** (Please refer to Statements C1 to C7)

28 When implementing the integrated sequence of lessons in the classroom, most teachers were convinced of the effectiveness of the suggested approach. The majority of teachers (about 90%) believed that pupils were more motivated to learn when teachers used the integrated approach and that the interaction generated by this approach promoted language learning. While more secondary teachers (63.6%) were confident of using this approach, fewer primary teachers (only 50%) felt confident enough. There was also a relatively large percentage of teachers (about 40%) who were uncertain about using this approach.

29 This uncertainty among some teachers could be due to their difficulty in adapting to a new teaching approach which made greater demands on the teacher's time and lesson preparation. For example, about 70% of the teachers felt that the approach took up too much time; 33.3% of the teachers felt that their pupils were bored by the sequence of lessons stretching for two weeks or more; and about 40% of teachers considered it difficult to teach grammar using this approach.

D Teacher Training and Materials Development (Please refer to Statement D1 to D4)

30 More teachers in the primary than secondary section (94.5% against 64.6%) felt that training in applying the approach was necessary. They thought that teachers who were trained in ACT or RSA courses would find it easier to use the integrated approach as advocated in the draft syllabus or to use the syllabus document itself. [RSA stands for Royal Society of Arts. This course for teachers leads to a Diploma in English Language teaching]. A fairly large proportion (45.5%) of secondary teachers, however, was rather uncertain, partly because of their ignorance of the existence of RSA courses or the PASSES programme in certain secondary schools. This was gleaned from their written comments in the completed questionnaires.

31 With regard to materials development, most teachers (100% in the primary section) indicated the need for the production of new materials based on the new syllabus. Many felt that existing course packages could not be easily adapted or exploited for use with the integrated approach, or were uncertain about how this could be done.

IMPLICATIONS

32 For this syllabus trialling, we had obtained much feedback from the the questionnaire completed by teachers. These data were complemented by classroom observations and discussions with the teachers, and in some cases, with pupils as well.

The feedback that was obtained has implications for teacher training and for revision of the draft syllabuses. With regard to the training of teachers, it was felt that teachers need to develop :

(a) a greater awareness of the integrative-interactive approach to language teaching/learning through the use of themes and topics

(b) the ability to exploit the various inventories for flexible lesson planning and integrated-interactive language use and learning. This includes the planning of appropriate and varied tasks/activities in relation to the cognitive/linguistic demands of the pupils and in relation to the language and language-related objectives that may be derived from the tasks/activities

(c) a deeper understanding of the role of grammar in language learning and how the Focus Lesson (explained in Chapter 2 of the syllabus) could be used for meaningful learning of form and accuracy

(d) a better understanding of the role of assessment in language learning

(e) better skills in time management to ensure that a planned sequence of lessons could be completed

(f) the ability to select, adapt and use learning materials to ensure that all lessons have sufficient interesting and varied audio-visual support

We shall be focussing on the above areas in the training workshops for all teachers teaching the lower blocks in the second semester of the year. In this connection, a training package has already been prepared.

33 Feedback from the trialling also pointed to the need to revise the draft syllabuses in a number of areas. For example, there was a need to simplify some linguistic terms and concepts in the draft syllabus to make the document more user-friendly. There was also a need to revise those chapters which contain the inventories to improve the ease of cross-referencing between chapters and to provide clearer guidelines in the chapter on Assessment.

CONCLUSIONS

34 It was gratifying to note that, in this syllabus trialling exercise, most teachers had cooperated and responded positively. From feedback given, the majority of teachers thought that the draft syllabuses were well organised and that the integrated approach advocated in the syllabuses, if properly implemented, would lead to more interesting and productive language learning for our pupils.

35 Relating to the objectives of the syllabus trialling, the findings had indicated that the teacher respondents were generally able to understand and interpret the contents of the syllabuses. With some guidance they had been able to plan a framework for an integrated sequence of lessons, using the various chapters and inventories in the syllabuses, and to select and adapt relevant materials for teaching and learning. We were also able to obtain useful information from the teachers for the revision of the draft syllabuses.

FINAL CONCLUSION (EPILOGUE)

36 Based on feedback and the findings of the trialling exercise, we made the necessary amendments and revisions to the draft syllabuses. This work was completed in the second semester of last year (1989), and the syllabuses have recently been distributed to all schools and training institutions. Given the necessary training in the use of the syllabuses, together with the availability of learning materials developed along the lines suggested in the syllabuses, we are confident that all teachers will be able to use the new syllabuses for more effective and efficient teaching and learning of English Language in the 1990s.

Appendix

QUESTIONNAIRE ON TRIALLING OF DRAFT EL SYLLABUSES (LOWER PRIMARY AND LOWER SECONDARY)

PART II

A THE DRAFT SYLLABUS AS A DOCUMENT

1 The chapters are well organized.

Response Level	Strongly Agree	Agree	Uncertain	Disagree	Strongly Disagree
Primary	11.1	72.2	0	16.7	0
Secondary	27.3	45.5	18.2	9.1	0

2 Cross-referencing is easy.

Response Level	Strongly Agree	Agree	Uncertain	Disagree	Strongly Disagree
Primary	5.6	61.1	5.6	22.2	5.6
Secondary	16.7	50.0	8.3	25.0	0

3 There is an overload of information.

Response Level	Strongly Agree	Agree	Uncertain	Disagree	Strongly Disagree
Primary	11.1	38.9	22.2	27.8	0
Secondary	0	36.4	27.3	36.4	0

4 The concepts are not well explained.

Response Level	Strongly Agree	Agree	Uncertain	Disagree	Strongly Disagree
Primary	0	38.9	16.7	44.4	0
Secondary	0	36.4	9.1	45.5	9.1

5 The language is too technical.

Response Level	Strongly Agree	Agree	Uncertain	Disagree	Strongly Disagree
Primary	27.8	50.0	0	22.2	0
Secondary	8.3	50.0	8.3	33.3	0

6 It provides clear guidelines for assessment of pupils' EL progress.

Response Level	Strongly Agree	Agree	Uncertain	Disagree	Strongly Disagree
Primary	5.6	27.8	0	33.3	33.3
Secondary	0	50.0	16.7	25.0	8.3

B <u>PLANNING AN INTEGRATED SEQUENCE OF LESSONS</u>

1 The syllabus can be used flexibly for lesson planning.

Response Level	Strongly Agree	Agree	Uncertain	Disagree	Strongly Disagree
Primary	16.7	83.3	0	0	0
Secondary	16.7	66.7	16.7	0	0

2 It can be used for planning an integrated sequence of lessons.

Response Level	Strongly Agree	Agree	Uncertain	Disagree	Strongly Disagree
Primary	11.1	88.9	0	0	0
Secondary	8.3	75.0	16.7	0	0

3 The examples in Chapter 2 help in the planning of an integrated sequence of lessons.

Response Level	Strongly Agree	Agree	Uncertain	Disagree	Strongly Disagree
Primary	22.2	77.8	0	0	0
Secondary	8.3	83.3	0	8.3	0

4 The Focus Wheel facilitates planning.

Response Level	Strongly Agree	Agree	Uncertain	Disagree	Strongly Disagree
Primary	11.1	88.9	0	0	0
Secondary	0	66.7	8.3	16.7	8.3

5 Planning takes up too much time.

Response Level	Strongly Agree	Agree	Uncertain	Disagree	Strongly Disagree
Primary	5C.0	44.4	0	5.6	0
Secondary	33.3	58.3	8.3	0	0

6 The approach suggested in the syllabus encourages the use of a variety of resources.

Response Level	Strongly Agree	Agree	Uncertain	Disagree	Strongly Disagree
Primary	33.3	66.7	0	0	0
Secondary	41.7	50.0	8.3	0	0

7 The suggested approach encourages a variety of teaching strategies.

Response Level	Strongly Agree	Agree	Uncertain	Disagree	Strongly Disagree
Primary	22.2	77.8	0	0	0
Secondary	16.7	75.0	8.3	0	0

8 The suggested approach encourages meaningful language use.

Response Level	Strongly Agree	Agree	Uncertain	Disagree	Strongly Disagree
Primary	11.1	66.7	16.7	5.6	0
Secondary	0	91.7	8.3	0	0

9 It is not possible to cover an adequate range of language items (eg grammar, functions, vocabulary, skills) in the lesson sequence.

Response Level	Strongly Agree	Agree	Uncertain	Disagree	Strongly Disagree
Primary	0	38.9	11.1	50.0	0
Secondary	16.7	25.0	8.3	50.0	0

10 The instructional objectives (IOs) are not easily derived from the tasks/activities.

Response / Level	Strongly Agree	Agree	Uncertain	Disagree	Strongly Disagree
Primary	5.6	66.7	22.2	5.6	0
Secondary	0	58.3	8.3	33.3	0

11 It is difficult to obtain suitable materials for teaching/learning.

Response / Level	Strongly Agree	Agree	Uncertain	Disagree	Strongly Disagree
Primary	11.1	66.7	11.1	11.1	0
Secondary	16.7	50.0	8.3	16.7	8.3

12 Assess the inventories in the syllabus using the 4-point scale below:-

1 2 3 4

Not at all ---------------------- Extremely

INVENTORY	Level	Comprehensive				Useful				Easy to Use			
		1	2	3	4	1	2	3	4	1	2	3	4
Themes	Pr	0	5.6	61.1	33.3	0	16.7	38.9	44.4	5.6	16.7	55.6	22.2
	Sec	0	16.7	41.7	41.7	0	25.0	33.3	41.7	0	33.3	33.3	33.3
Tasks and Activities	Pr	0	11.1	66.7	22.2	0	22.2	61.1	16.7	0	50.0	33.3	16.7
	Sec	0	16.7	75.0	8.3	0	16.7	75.0	8.3	0	33.3	66.7	0
Grammar	Pr	0	47.1	29.4	23.5	5.9	38.9	29.4	23.5	5.9	52.9	17.7	23.5
	Sec	8.3	16.7	41.7	33.3	8.3	41.7	33.3	16.7	16.7	16.7	50.0	16.7
Communicative Functions	Pr	0	44.4	16.7	38.9	5.6	33.3	38.9	22.2	11.1	33.3	38.9	16.7
	Sec	16.7	8.3	66.7	8.3	16.7	33.3	50.0	0	16.7	41.7	41.7	0
Spectrum of Skills	Pr	16.7	50.0	22.2	11.1	22.2	50.0	27.8	0	27.8	50.0	22.2	0
	Sec	8.3	16.7	66.7	8.3	16.7	41.7	33.3	8.3	25.0	33.3	41.7	0

C IMPLEMENTATION OF INTEGRATED SEQUENCE OF LESSONS

1 I am convinced of the effectiveness of the suggested approach.

Response Level	Strongly Agree	Agree	Uncertain	Disagree	Strongly Disagree
Primary	0	66.7	27.8	5.6	0
Secondary	8.3	75.0	16.7	0	0

2 I am confident about using this approach.

Response Level	Strongly Agree	Agree	Uncertain	Disagree	Strongly Disagree
Primary	0	50.0	44.4	5.6	0
Secondary	9.1	54.5	36.4	0	0

3 The pupils are motivated to learn by this approach.

Response Level	Strongly Agree	Agree	Uncertain	Disagree	Strongly Disagree
Primary	16.7	72.2	11.1	0	0
Secondary	0	90.9	9.1	0	0

4 The interactive method does not promote language learning.

Response Level	Strongly Agree	Agree	Uncertain	Disagree	Strongly Disagree
Primary	0	22.2	0	72.2	5.6
Secondary	0	18.2	27.3	36.4	18.2

5 This approach takes up too much time.

Response Level	Strongly Agree	Agree	Uncertain	Disagree	Strongly Disagree
Primary	16.7	50.0	11.1	22.2	0
Secondary	27.3	45.5	18.2	9.1	0

6 Pupils are bored by the integrated sequence of lessons stretched over 2 weeks or more.

Response Level	Strongly Agree	Agree	Uncertain	Disagree	Strongly Disagree
Primary	0	33.3	27.8	38.9	0
Secondary	0	33.3	25.0	33.3	8.3

7 It is difficult to teach grammar using this approach.

Response Level	Strongly Agree	Agree	Uncertain	Disagree	Strongly Disagree
Primary	5.6	33.3	16.7	44.4	0
Secondary	8.3	25.0	25.0	41.7	0

D TEACHER TRAINING AND MATERIALS DEVELOPMENT

1 A teacher needs to be trained to apply this approach.

Response Level	Strongly Agree	Agree	Uncertain	Disagree	Strongly Disagree
Primary	38.9	55.6	5.6	0	0
Secondary	8.3	58.3	16.7	8.3	8.3

2 Teachers trained in programmes/courses like ACT, RSA, PASSES will find it easy to use this approach/syllabus.

Response Level	Strongly Agree	Agree	Uncertain	Disagree	Strongly Disagree
Primary	16.7	66.7	16.7	0	0
Secondary	27.3	18.2	45.5	9.1	0

3 Course packages currently in use by the schools can be exploited easily for this approach.

Response Level	Strongly Agree	Agree	Uncertain	Disagree	Strongly Disagree
Primary	5.6	44.8	22.2	27.8	0
Secondary	0	25.0	41.7	25.0	8.3

4 New materials based on the syllabus will be most helpful.

Response / Level	Strongly Agree	Agree	Uncertain	Disagree	Strongly Disagree
Primary	88.9	11.1	0	0	0
Secondary	50.0	33.3	8.3	0	8.3

EAP PROGRAM EVALUATION IN AN ASIAN CONTEXT: A CASE FROM JAPAN

Mark Sawyer

INTRODUCTION

For numerous reasons, language program evaluation is seldom a smooth, straightforward process. To begin with, program administrators and curriculum developers often have only limited expertise on testing and evaluation. If they do possess the appropriate knowledge, it is still no easy matter to win the cooperation of all parties involved in order to effectively implement evaluation procedures; and even if all concerned are favorably disposed toward the endeavor, evaluations normally take place at the end of a program, when the administrators, teachers, and students need to redirect their thinking to the next term or program or project, or perhaps to a vacation. At the International University of Japan (IUJ), the English Language Program (ELP) is typical of language programs in general in that its evaluation process is subject to each of the above problems, but we have also had a certain amount of success in alleviating each of them.

The purpose of this paper is to describe certain aspects of the IUJ-ELP's evaluation process, in the hope that other language program evaluators can either benefit from our experiences or can provide us with some better ideas. My description will focus on ways we have tried to incorporate some very useful recent work that has been done on language program evaluation. I will start by showing how we have applied J.D. Brown's (1989) "systematic approach to curriculum improvement and maintenance," and then discuss our evaluation process in relation to Michael Long's (1984) ideas on "process and product" in evaluation. Finally, I will show how fruitful ideas for program evaluation can be generated by looking at the program from certain additional points of view: one is is as part of an "ecosystem" (Holliday and Cooke 1982), and the other in terms of "opportunity costs" (Swales 1989).

PROGRAM BACKGROUND

The IUJ- ELP was one of the first in a rapidly growing number of English for Academic Purposes (EAP) programs in Japan charged with preparing Japanese students to do English-medium academic studies. Whereas many of the other programs, such as the Japan programs of Temple University and Southern Illinois University, prepare students to do the main part of their studies in an English-speaking country, the entire curriculum at IUJ is undertaken on its campus in Niigata Prefecture. IUJ began with one academic program, a course in International Relations leading to an M.A. degree; in 1988 it established a second graduate school offering an M.B.A. in International Management. There are two main student populations: Japanese company employees who are sponsored by their employers to study for two years, and students from abroad who are attending on scholarship. These Japanese and non-Japanese populations are about equal in size (currently about 120 each), and the international students come from over thirty countries. In the ELP, we deal primarily with the Japanese students, because the international students tend to be stronger in English and in university-level study skills. All applicants to IUJ take TOEFL as part of the admissions procedure, the current range of our students being 450-670. The university is now in the process of establishing short-term non-degree courses, and there is talk of a third graduate school, dealing with International Development.

The ELP started out in 1983 with a full-time staff of one and four temporary instructors to run the first Intensive English Program (IEP) before the university formally opened. Now, seven years later, we have a full-time staff of seven faculty members, and an additional five to ten temporary instructors for our summer Intensive English Program, which has gradually expanded over the years to its current length of twelve weeks. In addition to the summer IEP, we offer two terms of credited EAP courses as a continuation of the work we began in the summer, but we consider the IEP to be the core of our curriculum and the time when students can clearly make progress. The IEP involves over seven hours of class per day, with a daily average of four hours of homework, and a full schedule of extracurricular activities. The IEP is currently divided into seven components, as shown in Figure 1.

Figure 1

A TYPICAL PROGRAM DAY

TIME	ACTIVITY
7.00-7.30	Morning Warm-up and Jogging
8.20-9.30	TEXT SKILLS I
9.40-10.50	TEXT SKILLS II
11.00-11.40	LANGUAGE LAB
11.50-12.30	ACCURACY DEVELOPMENT
12.30-1.30	Lunch
1.30-3.00	COMPUTER-ASSISTED INSTRUCTION
3.10-4.20	SEMINAR SKILLS I
4.30-5.40	SEMINAR SKILLS II
6.00-7.00	Dinner
7.30-8.30	Video Programs
8.30-9.30	Computer Room (assignments)

Although we have always been eager to gather information to help in developing the curriculum, applying our evaluation findings has always been an ambiguous undertaking, as there has never been a clear consensus on the goals of the university curriculum, and the addition of new programs has further confused the issue. Therefore, it is somewhat difficult for the ELP to judge clearly how successful it has been within its institutional context. This situation is further complicated by the fact that student success in dealing with the university's curriculum may not be closely related to success in terms of what their sponsors expect from them after graduation. Although the IUJ situation may be unique in some respects, there are ambiguities involved in the relationship between any language program and its wider institutional and societal context, and a sound evaluation process can go a long way toward resolving them.

THE "SYSTEMATIC" APPROACH

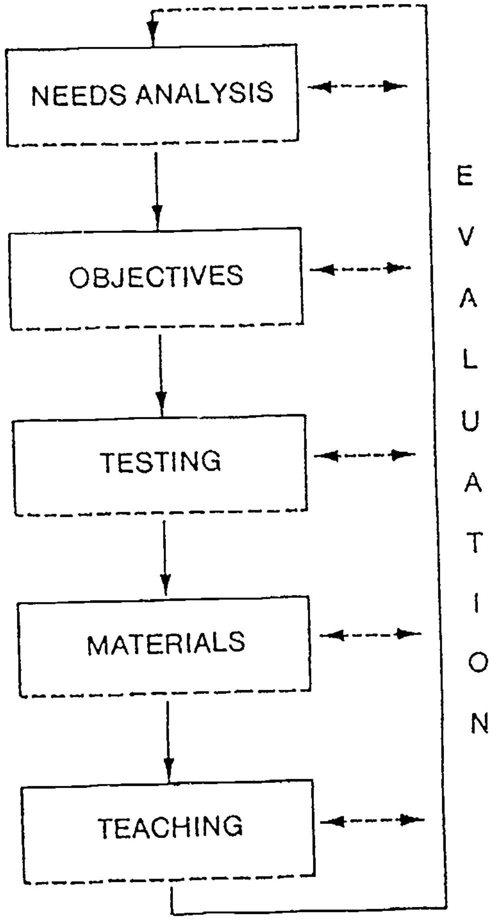
J. D. Brown defines evaluation as "the systematic collection and analysis of all relevant information necessary to promote the improvement of a curriculum and assess its effectiveness and efficiency, as well as participants' attitudes within the context of the particular institutions involved" (Brown, 1989:223). This definition may seem a bit awkward in its length, but the awkwardness itself serves to demonstrate that the concept of evaluation is sufficiently complex as to be very difficult to capture in one sentence. I would like to focus on two elements in the definition - systematicity and all relevant information - and show how they relate to the process of evaluation at IUJ.

Systematicity. Brown's use of "systematic" can be usefully interpreted in two ways. The first interpretation of "systematic" is that the data collection and analysis itself should be done systematically. This may seem rather obvious, but it is worth highlighting because this sense of "systematic" should not be taken to mean that data which is collected non-systematically can not be part of evaluation, but rather that the evaluators should work toward imposing as much systematicity as possible on data, some of which may appear very unsystematic at first. For example, comments made in the school cafeteria about the program by academic professors should not necessarily be disregarded, but if they are deemed to constitute useful information, then efforts should be made to gather a representative sample of them and record them consistently.

The second interpretation of "systematic" is that evaluation functions in a specific way as part of a system. A diagram of Brown's proposed model of evaluation will clarify this sense of "systematic."

Figure 2

Systematic approach for designing and maintaining language curriculum
(Brown, 1989:235)



As can be clearly seen, the system for designing curriculum goes through five stages from needs analysis to teaching, with evaluation providing input at each stage, and then the actual teaching provides input to revise the previous needs analysis. The most important thing to note about this model is that evaluation has a role at every stage, and is thus a constant process. It also should be understood that whereas curriculum design can follow a logical linear sequence from needs analysis to teaching, it need not, as evaluation occurring at any stage can have an influence on any other stage. This feature in the model gives recognition to the fact that evaluation is rarely a straightforward process.

All relevant information. The second important element of Brown's definition of evaluation to be focused on is the idea of using "all relevant information." Brown lists 24 data-gathering procedures which can be grouped into six categories, which can in turn be combined into one of two larger categories, according to the criterion of whether the evaluator is an "outsider looking in," or a "facilitator drawing information out" (Brown, 1989:233). This is a useful way to conceptualize the possibilities for evaluation procedures, but to describe the evaluation process at IUJ, it may be more useful to think in terms of types of information, and sources of of information. The types of information we use at IUJ include qualitative as well as quantitative, informal as well as formal, and information which is indirectly as well as directly relevant. The sources of information we use include student test scores, other program and university records, the students themselves, the ELP faculty, the academic faculties, administrators, alumni, sponsors, and other programs.

TYPES OF INFORMATION

Quantitative vs. qualitative. Concerning the use of both quantitative and qualitative information, there should not be much controversy: it is well accepted even in pure research that not everything can be quantified. However, there are other reasons for balancing out quantitative with qualitative data. Among the people who use evaluation reports to make decisions, there are doubtless many who find numbers most useful and/or impressive, but there are certainly others who are more or less innumerate, and there is a third group who like to see the presence of numbers to give them a sense of security that the evaluation is real, but in fact what they read is the qualitative part. Beretta (1989) cites a number of conflicting research findings on this issue. Another reason for not overemphasizing quantitative data is that such an emphasis limits the level of quality of the evaluation to the evaluator's level of statistical expertise . Of course it can be argued that it is the responsibility of a program evaluator be well-versed in research design and statistics, but the greater responsibility is in fact to evaluate the program as effectively as possible given the presently available resources.

Formal vs. informal. The use of informally gathered information may seem to contradict the need for systematicity, but there are good reasons to include it. The first is that people are often more candid in informal situations than in formal ones; it's quite possible that one can extract better information from some academic faculty members while chatting in the hallway than in even the most well-constructed questionnaire. The second reason is that the same faculty member is more likely to fill out that questionnaire conscientiously if he has recently spoken personally with one or more of the people responsible for it. Especially in a small educational community, and even more especially in Japan, informal contacts can often be considered a necessity before formal requests have any chance of being fulfilled.

Directly vs. indirectly relevant. Indirectly relevant information is that which does not seem to relate the program as currently conceived, but could play some role in the future. One example of the effective use of indirectly relevant data at IUJ can be seen during the period of the establishment of the School of International Management in affiliation with Dartmouth University's business school. At that time, the representatives from the Dartmouth side were enthusiastic about incorporating the "Rassias Method" of language teaching into the IUJ language curriculum. By researching this method in a timely way (in fact it is an audiolingually-oriented method which relies more or less on the teacher's manic behavior to keep the students attentive), and including a diplomatic critique of it in our program evaluation document, we were able to show them that we were of course familiar with the method, and that we were already incorporating its strong points into our program while avoiding its shortcomings. By doing this, we gained their support without having to explicitly accept or reject the method they were promoting.

SOURCES OF INFORMATION

Test scores. Student test scores are for us, as for most programs, the most compelling measure of success. We have been working hard to develop valid and reliable criterion-referenced tests to judge the effect of each component of our program, but we still have a long way to go. Moreover, even when we have achievement measures we are satisfied with, we will still be left with the equally difficult job of explaining them to non-specialists outside of the program whose decisions and attitudes may affect us.

Students. Evaluations of our courses by the students, although by no means an accurate measure of success, have nevertheless been by far the most useful source of feedback for us. We have found that once students realize that their opinion really is expected, and that it will be paid attention to, they often express their views quite thoughtfully and frankly. A lot of fine-tuning of our program has been possible due to student evaluations. One example is the level of intensity of our IEP. As could be seen in Figure 1, the program is very intensive, and this level must be maintained for twelve weeks. However, years of feedback from students concerning the intensity of the program have shown us that only a few consider the present level to be a hardship, and years when we lightened up the schedule a bit, many students felt somehow cheated.

ELP instructors . All instructors are asked to write out a brief formal evaluation after each course, but especiall,, .iteresting to us is the feedback we get from the instructors who come to teach in our summer Intensive English Program. While it is hard for the permanent staff to look objectively at the curriculum which they both designed and taught, our summer instructors are competent EAP teachers who have to implement an intensive curriculum without having had input into its content. Where weaknesses exist, the temporary instructors can generally be counted on to notice them.

Program and university records. University records provide us with data concerning learner performance outside the program, notably their grades in their academic courses and their final thesis grade. Since the students' degree of success in their academic programs is most often taken to be the ultimate measure of the success of an EAP Program, no amount of program-internal testing and evaluation will be sufficient. The problem with using course grades as a measure of success is that we know very little of the process by which professors arrived at those course grades. This has certainly been the case at IUJ, where course grades and level of English ability, or course grades and success in our IEP, have not correlated at all. Since nearly all students graduate with respectable grade point averages, we have not worried about this situation too much, but we are nevertheless vulnerable to criticism from the academic faculty.

Academic faculty. To pre-empt possible faculty criticism, it is very useful to find out as early as possible how professors view their students' language proficiency. The most systematic way to do this is through a questionnaire. This was done once during the first year of the program, and is now being done a second time (Uehara 1990). Regardless of the quality of the information from the questionnaire, it serves the purpose of showing the faculty that we care about our program, and if they do not use this convenient opportunity to voice their criticisms in a form where they can be constructive, those faculty members may be less likely to express their views later in a less constructive manner.

Administrators. Administrators play an extremely important role in the evaluation process, but in a different way than the parties mentioned so far. From time to time we can gather good information from them, but in general the crucial flow of information with administrators is in the other direction. The important thing is to present the results of valuation activities to them so that they will see good reason to support our program in the future.

Alumni. Alumni should be an excellent source of data about the effectiveness of the program, since they have the perspective to see how well the ELP and the IUJ curriculum prepared them for their new duties upon return to their companies. Regrettably, we are only now beginning to take on this job systematically, and have sent out a pilot questionnaire to 50 of our approximately 500 alumni. In terms of anecdotal data, however, we have had many productive informal conversations with alumni when they have returned to the campus for various events, and these conversations have led to a number of innovations.

Sponsors. It may also be worthwhile to see what the students' corporate sponsors think about the job we are doing, but it is difficult to obtain this information in any organized way, since we often do not even know what level in the sponsoring company we should address. Informal means of data collection may be useful, however. For example, when some of the company sponsors attend entrance and graduation ceremonies, our staff could take advantage of those opportunities to casually ask a few carefully chosen questions.

Other programs. Since there are many similarities in goals, methodology, and materials among EAP programs around the world, it seems irrational that there is not more communication among these programs. We have started to gather documentation from a number of programs in Japan and the United States, and we would be delighted to share materials with additional ones. Starting the year before last, we have also been sending one faculty member per year to teach in the Intensive English Program of the World Maritime

University in Sweden, and one of their faculty members will join our program this summer. This sort of staff exchange between programs with similar students and goals has already had significant benefits for both programs.

PROCESS AND PRODUCT EVALUATIONS

In his article on process and product evaluations, Michael Long focuses on process evaluation, defining it as "the systematic observation of classroom behavior with reference to the theory of (second) language development which underlies the program being evaluated" (Long 1984: 415). He points out that if product evaluation, carried out carefully with a true experimental design, and not subject to any threats to internal validity, shows a program to have arrived at the desired outcomes, we still cannot say that the curriculum or methodology caused those outcomes without examining actual classroom processes. Long does not discourage the use of product evaluations, but rather argues that both types are essential.

More specifically, what Long has in mind with process evaluation involves experimental research design with control groups and randomization, periodic video or audio recording, transcriptions, and careful examination of selected aspects of teacher and student behavior motivated by the SLA research literature, such as error correction, or ratio of referential to display questions. Although the rigor inherent in this approach is an ideal worthy of striving for, there are limitations in the approach Long advocates when applied to real programs, especially EAP programs.

The first limitation is that Long's approach does not seem to make a distinction between research and evaluation. Whereas research should lead to generalizable conclusions, evaluation must lead to specific decisions (Isaac & Michael, 1981). As Daniel Stufflebeam, a pioneer in modern educational evaluation, put it, "the purpose of evaluation is to improve, not to prove." Beretta (1986) argues along the same lines, claiming that Long's emphasis on internal validity shows that he does not recognize evaluation as the "applied research" it is. The kind of process evaluation outlined by Long could lead to answers to some fundamental questions of curriculum developers and teachers in the long run, but is not likely to provide the relevant answers needed in the short run.

A second limitation of Long's view of process evaluation is that it does not take into account the fact that EAP represents much more than language development. Therefore, it is not feasible to evaluate the program in terms of any one theory. At the very least, we need an SLA theory and an academic skill learning theory.

A third and related limitation is the practical one of being constrained by expertise, time, and motivation. The rigorous procedure outlined by Long would require considerable amounts of each. Without the full cooperation of a good-sized staff of experts who buy into the same theories and and accept the same research priorities, it would be difficult to see such an evaluation process through to completion. On the other hand, this limitation does not imply that the disciplined process evaluation is not a worthy ideal to pursue, but merely shows that our expectations cannot be too high at first, while the first two limitations serve mostly to caution that other concerns need to be balanced against those that Long emphasizes.

IUJ application of process evaluation. While we cannot claim to be carrying out process evaluations according to Long's criteria, we do take steps to ensure that our curriculum is carried out as devised. For our summer IEP, we start by sending out course syllabi and teaching materials to temporary instructors before they arrive. Upon arrival, we hold pre-service component orientation meetings, and component coordination meetings continue on an ongoing basis. The director observes every section of every component at least once. Additionally, all instructors keep a cumulative lesson plan notebook on their desks, so that instructors of the same (or different) components can compare notes, and the component coordinators and program director can see how well the curriculum is being followed.

Midterm student evaluations also provide clues to corroborate that the curriculum is really being followed, and that is is worthy of following. An additional feature in our next summer program will be student committees for each component. These committees will consist of a representative from each section, and will meet once a week to discuss with the component coordinator and the program director differences in their experiences.

PROCESS AND PRODUCT OF EVALUATION

The experience of the IUJ-ELP in relation to program evaluation can be best captured by borrowing the terms process and product, and using them in a completely different way from the previous section. The process of evaluation will roughly correspond to the formative aspects of evaluation, and the products of evaluation will primarily refer to the summative aspects of evaluation.

PROCESS

The elements of time, expertise, and motivation have been previously mentioned as constraints affecting the success of program evaluation. What is needed is an evaluation process which can maximize the amounts of these elements available. To increase the amount of time available, it is necessary for evaluation to become a high priority activity, and for as many people as possible to be share the work. To increase the amount of expertise available, it is necessary to provide direction and encouragement, and to delimit the areas of expertise necessary for each staff member. Finally, to increase motivation, everyone involved (especially, but not exclusively, the program staff) needs to understand how they themselves can benefit from the evaluation. Of course, it is very easy to see how these elements overlap and affect each other.

Time. In the IUJ-ELP, each faculty member has a responsibility for a certain component of the curriculum, a responsibility which includes evaluation. Dividing the work up among seven faculty members makes the workload bearable, and the director is freed to take responsibility for promoting consistency in the evaluation procedures across components, and coherence in the curriculum as a whole.

Expertise. With regard to expertise, we can generally assume sufficient knowledge in the area of curriculum development when we hire our faculty, but this is certainly not the case with regard to the evaluation of that curriculum. We started a campaign two years ago to increase knowledge on topics related to evaluation, with more experienced staff giving faculty colloquia on topics such as criterion-referenced testing, behavioral objectives, and questionnaire design. Starting this year, we are also delegating stages in Brown's (1989) model of curriculum development to the faculty , i.e. one member will be primarily responsible for organizing and serving as a resource person for needs analysis, another for objectives, etc.

Motivation. Motivation for evaluation is generated by giving the staff good reason to participate fully in the process. In the IUJ-ELP, everyone realizes that the curriculum they developed will get even better as a result of the evaluation process, and they also know that the degree to which they get useful information depends on the amount of care they put into designing their evaluations. Likewise, the degree to which their students' progress will be evident depends on the quality of the tests the instructors devise.

Another reason to get involved in the process is the opportunity for professional development. At IUJ this is especially easy, because faculty promotion is based on a point system that recognizes faculty colloquia and conference presentations. The result is that individual faculty members explore some new area which is most often related to curriculum or evaluation, provide new knowledge to others through a faculty colloquium, get feedback from the faculty, develop it further into a conference presentation, get more feedback, and then apply the by then well-developed ideas back to their area of the curriculum or evaluation process. The recent papers by Hayes (1990) and Uehara (1990), as well as the present paper, are all parts of this process.

An additional point to made about motivation is that the evaluation process at IUJ is essentially non-threatening. One reason this is possible is that our emphasis is always on curriculum rather than teacher evaluation. Debriefings after class observations by the director center around variations in approaches to implementing the curriculum, and although instructors may add questions to student evaluation forms specifically relating to their teaching technique, the program-wide questions focus principally on syllabus and materials . Of course, anticipating the results of each set of student course evaluations brings some anxiety, but since evaluation is now an integrated part of the program culture, this anxiety becomes a routine aspect of the teaching experience, and translates into incentive to further improve the course. The IUJ-ELP currently de-emphasizes teacher evaluation in order to maintain total commitment of the teaching staff to the evaluation process and to the professional development that goes along with it. However, as the curriculum gets more and more finely tuned, and commitment to the process is less of an issue, it is easy to foresee a time when teacher evaluation may be highlighted more.

PRODUCT

Although the process of evaluation itself serves the formative purpose within the program of providing direction for future improvement, it is usually the case that administrators and other groups in decision-making positions require some product with which to assess the effectiveness and efficiency of the program. In the case of the IUJ-ELP, we have been fortunate in having very little direct outside interference, but as we grow and demand more resources, members of the Policy Shaping Community, or PSC (see Beretta, 1990) have begun to take more interest in our use of those resources.

Following the principle of division of labor and expertise elaborated above, we now produce an annual Intensive English Program Final Report, with each component coordinator writing the sections relevant to that component. The emphasis is on readability, in terms of vocabulary (no SLA, or evaluation, jargon), balance between qualitative and quantitative data, and adjustable length. The report starts with a one-page very broad overview by the director, then one-page component overviews by each of the coordinators, then one-page descriptions of the criterion-referenced tests used, then summaries of student evaluations, and so on, gradually moving into charts and graphs of student performance and finally some of the relevant raw data. Our intent is that PSC members will read as little or as much as they wish, but come away with an overall picture of the nature and success of the program, and the level of effort and professionalism that we put into it. We do not try to hide problems, but unless they are problems that the particular reader can potentially assist us in solving, we mention them straightforwardly but briefly, with an intended solution immediately following.

We are also in the process of producing an ongoing program document, again divided into many short sections, in which we are compiling general program information, a description of the curriculum development process, a rationale for our curriculum, brief summaries of student performance and evaluation data for year-by-year comparison, etc. This report is intended to be useful to new instructors and to people from other EAP programs. It also should prove useful to ourselves, in serving as a reason to step back from time to time and look at the program as a whole.

THE ECOLOGICAL APPROACH

It is relatively easy to obtain understanding and forge a consensus among the IUJ-ELP staff on most program-internal matters, but on issues involving parties both inside and outside the program, this harmony is not always the case. In Japan, as in any foreign country, the hosts' ways of dealing with things sometimes do not seem to make sense. In these situations, the ecological approach proposed by Holliday and Cooke (1982) provides a very useful metaphor. Holliday and Cooke see language programs as existing within" a milieu of attitudes and expectations of all the parties involved...we treat this milieu as an ecosystem within which we have to work. The novelty of our approach lies in the practical implications of this view: the need to accords rights of co-existence to <u>all</u> the competing but interdependent elements of the system, and to work with the system, rather than against or in spite of it, to the greatest extent possible" (Holliday and Cooke 1982:126). The goal of the program is thus to make the best use of local features, both promising and unpromising, so that the long-term viability of the project or program can be assured. In such a situation one crucial function of evaluation is to gather the relevant data on local features and the interests of all the parties functioning within the ecosystem. This approach to program design is especially appropriate for Japan, where the rule is consensus-type decision-making, in which the needs of all relevant parties are typically weighed into a final decision. Attempts to impose different modes of decision-making almost invariably result in frustration.

OPPORTUNITY COST CONSIDERATIONS

John Swales (1989) discusses the importance in curriculum development of understanding the decision-making process that goes into it. He suggests the usefulness of applying the economic concept opportunity cost , defined as "real or full costs, taking into consideration the deficits created by the forced abandonment of other alternatives" (Swales 1989: 82). In other words, every decision involves sacrificing options associated with a different course of action. For example, deciding to adopt Textbook B eliminates the possibility of enjoying the advantages of your current Textbook A, or other textbooks C, D, or E, or no fixed textbook at all. In a sense, focusing on opportunity costs can be seen as conservatizing since it emphasizes the negative implications of any decision, but in fact it equally involves evaluating the costs of not innovating.

A clear-cut case of opportunity cost considerations influencing an IUJ-ELP decision occurred in planning for our 1989 IEP, in which we had to decide whether or not to introduce a new program component focusing explicitly on grammatical accuracy. Although the program director and some of the senior staff felt that such a course would be pedagogically unsound, some of the newer staff members were totally convinced that this type of course was necessary. Furthermore, there was an abundance of anecdotal evidence that most of our students would appreciate such a course. To reject the course meant that the staff members supporting it would then be less committed to the program as a whole, and during a long intensive program, the probability was high that students would somehow find out about this course that had been denied them. Thus, recognizing these opportunity costs, we initiated the course. As a result, the staff members in favor of it worked hard on the curriculum to make it work, students liked it, their post-test scores were encouraging, and even the skeptical instructors who were asked to teach a section of it saw some value in it.

The above case was one in which opportunity cost considerations aided in making a wise curriculum decision. These considerations could be even more applicable when making decisions concerning the program's relationship with its institutional environment, for example, in deciding whether the program should offer new short-term courses, or whether it should become independent of the parent institution. In such situations, Swales (1989) argues that the ecological approach's emphasis on understanding how the system works is not enough, and that the concept of opportunity cost can provide more specific guidance in making the right strategic decisions.

CONCLUSION

Although the literature on language program evaluation is still in its incipient stages, it has already offered some sound and usable ideas. From J.D. Brown (1989) we have taken an overall framework for curriculum development; from Michael Long (1984) we have realized the importance of investigating the process that we assume has produced our program results, and of working toward greater scientific discipline in our evaluation efforts; from Holliday and Cooke (1982) we have gained an appreciation of the need to understand our role as a part of a complex larger system; and from John Swales (1989) we have received an approach to applying the information we have gathered toward making strategic decisions.

Although the IUJ-ELP still suffers to some degree from most of the recurrent problems of program design enumerated by Swales (1989:86), we feel that through our evaluation process, we have been able to make each problem much less severe than it would be otherwise; we have also been able to determine with increasing accuracy which problems we should vigorously continue to try to solve, and which we should simply accept as unfavorable "local features," which may prove in the future to have "positive and exploitable aspects" (Holliday and Cooke, 1982:137).

If there is anything that other programs can learn from us, it is the benefits of making the process and products of evaluation an integral part of program culture. When the evaluation process becomes second nature to a program staff, any weaknesses in the procedures themselves will be overshadowed by the commitment of the staff to the program and the virtual guarantee of ongoing improvement in the future. Furthermore, when the regular reporting of evaluation findings becomes an established practice, in a way that adequately considers the needs of each audience, the chances of obtaining enhanced institutional support are sure to increase. Finally, it seems reasonable to assume that there are many things that EAP programs around the world can learn from each other; sound evaluation practices of course contribute to having something to say, but by far the most important first step is simply communication. We at the International University of Japan look forward to sharing ideas and experiences with other EAP programs.

REFERENCES

Barrett, R. (ed.). 1982: The Administration of Intensive English Language Programs. Washington, D.C.: National Association of Foreign Student Affairs.

Brown, J. 1989. Language program evaluation: a synthesis of existing possibilities. In The Second Language Curriculum, R. Johnson (ed.), 222-241. Cambridge: Cambridge University Press.

Beretta, A. 1986a. Toward a methodology of ESL program evaluation. TESOL Quarterly 20(1):144-55.

Beretta, A. 1986b. A case for field experimentation in program evaluation. Language Leaming 36(3): 295-310.

Beretta, A. 1989. The program evaluator: the ESL researcher without portfolio. Applied Linguistics 16(1):1-15.

Hayes, T. (1990). Progressive evaluation of academic writing of graduate students at the International University of Japan. Paper presented at the 1990 RELC Regional Seminar, Singapore.

Holliday, A. and T. Cooke. 1982. An ecological approach to ESP. In Issues in ESP, A. Waters, (ed.), 123-43. Oxford: Pergamon.

Isaac, S. and W. Michael. 1981. Handbook in Research and Evaluation. San Diego, CA: Edits Publishers.

Johnson, R. 1989. The Second Language Curriculum. Cambridge: Cambridge University Press.

Long, M. 1984. Process and Product in ESL program evaluation. TESOL Quarterly 18(3): 409-25.

Pennington, M. and A. Young. 1989. Approaches to faculty evaluation for ESL. TESOL Quarterly 23(3): 619-46.

Pratt, D. 1980. Curriculum Design and Development. New York: Harcourt Brace Jovanovich.

Swales, J. 1989. Service English programme design and opportunity cost. In The Second Language Curriculum, R. Johnson (ed.), 79-90. Cambridge: Cambridge University Press.

Uehara, Randal. (1990). Language program evaluation: a unique approach. Paper presented at the 1990 RELC Regional Seminar, Singapore.

White, R. 1988. The ELT Curriculum. Oxford: Basil Blackwell.

LIST OF CONTRIBUTORS

Prof Margaret Des Brisay
Project Director
Canadian Test of English for
Scholars and Trainees
Second Language Institute
University of Ottawa
600 King Edward
Ottawa K1N 6NS
CANADA

Mr David Crabbe
English Language Institute
Victoria University of Wellington
P O Box 600
Wellington
NEW ZEALAND

Mr Ali Abdul Ghani
Assistant Director
Language Unit, Schools Division
Ministry of Education
Level 5, Block J (South)
Pusat Bandar Damansara
50604 Kuala Lumpur
MALAYSIA

Mr Goh Soon Guan
Specialist Inspector
English Language 6
Curriculum Planning Division
Languages and Library Development Branch
Ministry of Education
Environment Building, 8th Floor
40 Scotts Road
SINGAPORE 0922

Ms Elvira C Fonacier
Assistant Professor
Languages Department
De La Salle University
2401 Taft Avenue
Manila
PHILIPPINES

Prof Harry L Gradman
Professor of Linguistics
Department of Linguistics
Memorial Hall 313
Indiana University
Bloomington, Indiana 47405
U S A

Dr Edith Hanania
Centre for English Language Training
Memorial Hall 313
Indiana University
Bloomington, Indiana 47405
U S A

Mr Brian Hunt
English Language Adviser
Schools Division
Ministry of Education
Level 5, Block J (South)
Pusat Bandar Damansara
50604 Kuala Lumpur
MALAYSIA

Dr Ronald Mackay
Associate Professor of Applied Linguistics
Centre for the Teaching of English as a Second Language
Concordia University
Montreal
CANADA

Dr Alastair L McGregor
Principal Lecturer, TESOL
Western Australian College of Advanced Education
2 Bradford Street
Mount Lawley, W A 60250
AUSTRALIA

Dr Ma Flor E Mejorada
Director
Writing Laboratory
Languages Department
De La Salle University
2401 Taft Avenue
Manila
PHILIPPINES

Mr Dermot Murphy
Senior Lecturer
St Mary's College, Twickenham
12 Clapham Common North Side
London SW4 ORF
UNITED KINGDOM

Dr David Nunan
Associate Director
National Centre for English
Language Teaching and Research
Macquarie University
North Ryde, N S W 2109
AUSTRALIA

Dr Adrian Palmer
Associate Professor
Department of English
The University of Utah
OSH-341
Salt Lake City, Utah 84112
U S A

Ms Doreen Ready
Head
Testing and Research Support Services
Second Language Institute
University of Ottawa
600 King Edward
Ottawa K1N 6NS
CANADA

Mr John Roberts
University of Reading
Whiteknights
P O Box 218
Reading RG6 2AA
UNITED KINGDOM

Prof Mark Sawyer
Associate Professor/ELP Director
International University of Japan
Yamato-machi, Minami
Uonuma-gun
Niigata-ken 949-72
JAPAN

Dr C J Weir
EFL Lecturer
CALS Testing and Evaluation Unit
University of Reading
Whiteknights
P O Box 218
Reading RG6 2AA
UNITED KINGDOM

Printed by Sin Sin Lithographers Pte Ltd.